Contents

Preface

Date: 01 May 2023

Book Description

Loss Data Analytics is an interactive, online, freely available text.

- The online version contains many interactive objects (quizzes, computer demonstrations, interactive graphs, video, and the like) to promote *deeper learning*.
- A subset of the book is available for *offline reading* in pdf and EPUB formats.
- The online text will be available in multiple languages to promote access to a *worldwide audience*.

What will success look like?

The online text will be freely available to a worldwide audience. The online version will contain many interactive objects (quizzes, computer demonstrations, interactive graphs, video, and the like) to promote deeper learning. Moreover, a subset of the book will be available in pdf format for low-cost printing. The online text will be available in multiple languages to promote access to a worldwide audience.

How will the text be used?

This book will be useful in actuarial curricula worldwide. It will cover the loss data learning objectives of the major actuarial organizations. Thus, it will be suitable for classroom use at universities as well as for use by independent learners seeking to pass professional actuarial examinations. Moreover, the text will also be useful for the continuing professional development of actuaries and other professionals in insurance and related financial risk management industries.

Why is this good for the profession?

An online text is a type of open educational resource (OER). One important benefit of an OER is that it equalizes access to knowledge, thus permitting a broader community to learn about the actuarial profession. Moreover, it has the capacity to engage viewers through active learning that deepens the learning process, producing analysts more capable of solid actuarial work.

Why is this good for students and teachers and others involved in the learning process? Cost is often cited as an important factor for students and teachers in textbook selection (see a recent post on the $400 textbook). Students will also appreciate the ability to "carry the book around" on their mobile devices.

Why loss data analytics?

The intent is that this type of resource will eventually permeate throughout the actuarial curriculum. Given the dramatic changes in the way that actuaries treat data, loss data seems like a natural place to start. The idea behind the name *loss data analytics* is to integrate classical loss data models from applied probability with modern analytic tools. In particular, we recognize that big data (including social media and usage based insurance) are here to stay and that high speed computation is readily available.

Project Goal

The project goal is to have the actuarial community author our textbooks in a collaborative fashion. To get involved, please visit our Open Actuarial Textbooks Project Site.

Acknowledgements

Edward Frees acknowledges the John and Anne Oros Distinguished Chair for Inspired Learning in Business which provided seed money to support the project. Frees and his Wisconsin colleagues also acknowledge a Society of Actuaries Center of Excellence Grant that provided funding to support work in dependence modeling and health initiatives. Wisconsin also provided an education innovation grant that provided partial support for the many students who have worked on this project.

We acknowledge the Society of Actuaries for permission to use problems from their examinations.

We thank Rob Hyndman, Monash University, for allowing us to use his excellent style files to produce the online version of the book.

We thank Yihui Xie and his colleagues at Rstudio for the R bookdown package that allows us to produce this book.

We also wish to acknowledge the support and sponsorship of the International Association of Black Actuaries in our joint efforts to provide actuarial educational content to all.

Contributors

The project goal is to have the actuarial community author our textbooks in a collaborative fashion. The following contributors have taken a leadership role in developing *Loss Data Analytics*.

- **Zeinab Amin** is a Professor at the Department of Mathematics and Actuarial Science and Associate Provost for Assessment and Accreditation at the American University in Cairo (AUC). Amin holds a PhD in Statistics and is an Associate of the Society of Actuaries. Amin is the recipient of the 2016 Excellence in Academic Service Award and the 2009 Excellence in Teaching Award from AUC. Amin has designed and taught a variety of statistics and actuarial science courses. Amin's current area of research includes quantitative risk assessment, reliability assessment, general statistical modelling, and Bayesian statistics.

- **Katrien Antonio**, KU Leuven

- **Jean-François Bégin** is an Assistant Professor in the Department of Statistics and Actuarial Science at Simon Fraser University in British Columbia, Canada. Bégin holds a PhD in Financial Engineering from HEC Montréal, Canada, and is a Fellow of the Society of Actuaries and of the Canadian Institute of Actuaries. His current research interests include financial modelling, financial econometrics, Bayesian statistics, filtering methods, credit risk, option pricing, and pension economics. Bégin has designed and taught a variety of actuarial finance and actuarial communication courses.

- **Jan Beirlant**, KU Leuven

- **Arthur Charpentier** is a professor in the Department of Mathematics at the Université du Québec á Montréal. Prior to that, he worked at a large general insurance company in Hong Kong, China, and the French Federation of Insurers in Paris, France. He received a MS on mathematical economics at Université Paris Dauphine and a MS in actuarial science at ENSAE (National School of Statistics) in Paris, and a PhD degree from KU Leuven, Belgium. His research interests include econometrics, applied probability and actuarial science. He has published several books (the most recent one on *Computational Actuarial Science with R*, CRC) and papers on a variety of topics. He is a Fellow of the French Institute of

Actuaries, and was in charge of the 'Data Science for Actuaries' program from 2015 to 2018.

- **Curtis Gary Dean** is the Lincoln Financial Distinguished Professor of Actuarial Science at Ball State University. He is a Fellow of the Casualty Actuarial Society and a CFA charterholder. He has extensive practical experience as an actuary at American States Insurance, SAFECO, and Travelers. He has served the CAS and actuarial profession as chair of the Examination Committee, first editor-in-chief for *Variance: Advancing the Science of Risk*, and as a member of the Board of Directors and the Executive Council. He contributed a chapter to *Predictive Modeling Applications in Actuarial Science* published by Cambridge University Press.

- **Edward W. (Jed) Frees** is an emeritus professor, formerly the Hickman-Larson Chair of Actuarial Science at the University of Wisconsin-Madison. He is a Fellow of both the Society of Actuaries and the American Statistical Association. He has published extensively (a four-time winner of the Halmstad and Prize for best paper published in the actuarial literature) and has written three books. He also is a co-editor of the two-volume series *Predictive Modeling Applications in Actuarial Science* published by Cambridge University Press.

- **Guojun Gan** is an associate professor in the Department of Mathematics at the University of Connecticut, where he has been since August 2014. Prior to that, he worked at a large life insurance company in Toronto, Canada for six years. He received a BS degree from Jilin University, Changchun, China, in 2001 and MS and PhD degrees from York University, Toronto, Canada, in 2003 and 2007, respectively. His research interests include data mining and actuarial science. He has published several books and papers on a variety of topics, including data clustering, variable annuity, mathematical finance, applied statistics, and VBA programming.

- **Lisa Gao** is a PhD candidate in the Risk and Insurance department at the University of Wisconsin-Madison. She holds a BMath in Actuarial Science and Statistics from the University of Waterloo and is an Associate of the Society of Actuaries.

- **José Garrido**, Concordia University

- **Lei (Larry) Hua** is an Associate Professor of Actuarial Science at Northern Illinois University. He earned a PhD degree in Statistics from the University of British Columbia. He is an Associate of the Society of Actuaries. His research work focuses on multivariate dependence modeling for non-Gaussian phenomena and innovative applications for financial and insurance industries.

- **Noriszura Ismail** is a Professor and Head of Actuarial Science Program, Universiti Kebangsaan Malaysia (UKM). She specializes in Risk Modelling

and Applied Statistics. She obtained her BSc and MSc (Actuarial Science) in 1991 and 1993 from University of Iowa, and her PhD (Statistics) in 2007 from UKM. She also passed several papers from Society of Actuaries in 1994. She has received several research grants from Ministry of Higher Education Malaysia (MOHE) and UKM, totaling about MYR1.8 million. She has successfully supervised and co-supervised several PhD students (13 completed and 11 on-going). She currently has about 180 publications, consisting of 88 journals and 95 proceedings.

- **Joseph H.T. Kim**, Ph.D., FSA, CERA, is Associate Professor of Applied Statistics at Yonsei University, Seoul, Korea. He holds a Ph.D. degree in Actuarial Science from the University of Waterloo, at which he taught as Assistant Professor. He also worked in the life insurance industry. He has published papers in *Insurance Mathematics and Economics*, *Journal of Risk and Insurance*, *Journal of Banking and Finance*, *ASTIN Bulletin*, and *North American Actuarial Journal*, among others.

- **Nii-Armah Okine** is an assistant professor at the Mathematical Sciences Department at Appalachian State University. He holds a Ph.D. in Business (Actuarial Science) from the University of Wisconsin - Madison and obtained his master's degree in Actuarial science from Illinois State University. His research interest includes micro-level reserving, joint longitudinal-survival modeling, dependence modeling, micro-insurance, and machine learning.

- **Emine Selin Sarıdaş** is a doctoral candidate in the Statistics department of Mimar Sinan University. She holds a bachelor degree in Actuarial Science with a minor in Economics and a master degree in Actuarial Science from Hacettepe University. Her research interest includes dependence modeling, regression, loss models and life contingencies.

- **Peng Shi** is an associate professor in the Risk and Insurance Department at the Wisconsin School of Business. He is also the Charles & Laura Albright Professor in Business and Finance. Professor Shi is an Associate of the Casualty Actuarial Society (ACAS) and a Fellow of the Society of Actuaries (FSA). He received a Ph.D. in actuarial science from the University of Wisconsin-Madison. His research interests are problems at the intersection of insurance and statistics. He has won several research awards, including the Charles A. Hachemeister Prize, the Ronald Bornhuetter Loss Reserve Prize, and the American Risk and Insurance Association Prize.

- **Nariankadu D. Shyamalkumar (Shyamal)** is an associate professor in the Department of Statistics and Actuarial Science at The University of Iowa. He is an Associate of the Society of Actuaries, and has volunteered in various elected and non-elected roles within the SoA. Having a broad theoretical interest as well as interest in computing, he has published in prominent actuarial, computer science, probability theory, and statistical journals. Moreover, he has worked in the financial industry, and since

then served as an independent consultant to the insurance industry. He has experience educating actuaries in both Mexico and the US, serving in the roles of directing an undergraduate program, and as a graduate adviser for both masters and doctoral students.

- **Jianxi Su** is an Assistant Professor at the Department of Statistics at Purdue University. He is the Associate Director of Purdue's Actuarial Science. Prior to joining Purdue in 2016, he completed the PhD at York University (2012-2015). He obtained the Fellow of the Society of Actuaries (FSA) in 2017. His research expertise are in dependence modelling, risk management, and pricing. During the PhD candidature, Jianxi also worked as a research associate at the Model Validation and ORSA Implementation team of Sun Life Financial (Toronto office).

- **Chong It Tan** is a senior lecturer at Macquarie University in Australia, where he has served as the undergraduate actuarial program director since 2018. He obtained his PhD in 2015 from Nanyang Technological University in Singapore. He is a fully qualified actuary, holding the credentials from both the US Society of Actuaries and Australian Actuaries Institute. His major research interests are mortality modelling, longevity risk management and bonus-malus systems.

- **Tim Verdonck** is associate professor at the University of Antwerp. He has a degree in Mathematics and a PhD in Science: Mathematics, obtained at the University of Antwerp. During his PhD he successfully took the Master in Insurance and the Master in Financial and Actuarial Engineering, both at KU Leuven. His research focuses on the adaptation and application of robust statistical methods for insurance and finance data.

- **Krupa Viswanathan** is an Associate Professor in the Risk, Insurance and Healthcare Management Department in the Fox School of Business, Temple University. She is an Associate of the Society of Actuaries. She teaches courses in Actuarial Science and Risk Management at the undergraduate and graduate levels. Her research interests include corporate governance of insurance companies, capital management, and sentiment analysis. She received her Ph.D. from The Wharton School of the University of Pennsylvania.

Reviewers

Our goal is to have the actuarial community author our textbooks in a collaborative fashion. Part of the writing process involves many reviewers who generously donated their time to help make this book better. They are:

- Yair Babab
- Chunsheng Ban, Ohio State University
- Vytaras Brazauskas, University of Wisconsin - Milwaukee

- Yvonne Chueh, Central Washington University
- Chun Yong Chew, Universiti Tunku Abdul Rahman (UTAR)
- Eren Dodd, University of Southampton
- Gordon Enderle, University of Wisconsin - Madison
- Rob Erhardt, Wake Forest University
- Runhun Feng, University of Illinois
- Brian Hartman, Brigham Young University
- Liang (Jason) Hong, University of Texas at Dallas
- Fei Huang, Australian National University
- Hirokazu (Iwahiro) Iwasawa
- Himchan Jeong, University of Connecticut
- Min Ji, Towson University
- Paul Herbert Johnson, University of Wisconsin - Madison
- Dalia Khalil, Cairo University
- Samuel Kolins, Lebonan Valley College
- Andrew Kwon-Nakamura, Zurich North America
- Ambrose Lo, University of Iowa
- Mark Maxwell, University of Texas at Austin
- Tatjana Miljkovic, Miami University
- Bell Ouelega, American University in Cairo
- Zhiyu (Frank) Quan, University of Connecticut
- Jiandong Ren, Western University
- Rajesh V. Sahasrabuddhe, Oliver Wyman
- Sherly Paola Alfonso Sanchez, Universidad Nacional de Colombia
- Ranee Thiagarajah, Illinois State University
- Ping Wang, Saint Johns University
- Chengguo Weng, University of Waterloo
- Toby White, Drake University
- Michelle Xia, Northern Illinois University
- Di (Cindy) Xu, University of Nebraska - Lincoln
- Lina Xu, Columbia University
- Lu Yang, University of Amsterdam
- Jorge Yslas, University of Copenhagen
- Jeffrey Zheng, Temple University
- Hongjuan Zhou, Arizona State University

Other Collaborators

- Alyaa Nuval Binti Othman, Aisha Nuval Binti Othman, and Khairina (Rina) Binti Ibraham were three of many students at the Univeristy of Wiscinson-Madison that helped with the text over the years.
- Maggie Lee, Macquarie University, and Anh Vu (then at University of New South Wales) contributed the end of the section quizzes.
- Jeffrey Zheng, Temple University, Lu Yang (University of Amsterdam), and Paul Johnson, University of Wisconsin-Madison, led the work on the glossary.

Version

- This is **Version 1.1**, August 2020. Edited by Edward (Jed) Frees and Paul Johnson.
- Version 1.0, January 2020, was edited by Edward (Jed) Frees.

You can also access pdf and epub (current and older) versions of the text in our Offline versions of the text.

For our Readers

We hope that you find this book worthwhile and even enjoyable. For your convenience, at our Github Landing site (https://openacttexts.github.io/), you will find links to the book that you can (freely) download for offline reading, including a pdf version (for Adobe Acrobat) and an EPUB version suitable for mobile devices. Data for running our examples are available at the same site.

In developing this book, we are emphasizing the online version that has lots of great features such as a glossary, code and solutions to examples that you can be revealed interactively. For example, you will find that the statistical code is hidden and can only be seen by clicking on terms such as

We hide the code because we don't want to insist that you use the `R` statistical software (although we like it). Still, we encourage you to try some statistical code as you read the book – we have opted to make it easy to learn `R` as you go. We have set up a separate R Code for Loss Data Analytics site to explain more of the details of the code.

Like any book, we have a set of notations and conventions. It will probably save you time if you regularly visit our Appendix Chapter **??** to get used to ours.

Freely available, interactive textbooks represent a new venture in actuarial education and we need your input. Although a lot of effort has gone into the development, we expect hiccoughs. Please let your instructor know about opportunities for improvement, write us through our project site, or contact chapter contributors directly with suggested improvements.

Chapter 1

Bayesian Inference and Modeling

Chapter Preview. Up to this point in the book, we have focused almost exclusively on the frequentist approach to estimate our various loss distribution parameters. In this chapter, we switch gears and discuss a different paradigm: Bayesianism. These approaches are different as Bayesian and frequentist statisticians disagree on the source of the uncertainty: Bayesian statistics assumes that the observed data sample is fixed and that model parameters are random, whereas frequentism considers the opposite (i.e., a random sample but fixed—yet unknown—model parameters).

In this chapter, we introduce Bayesian inference and modeling with a particular focus on loss data analytics. We begin in Section 1.1 by explaining the basics of Bayesian statistics: we compare it to frequentism and provide some historical context for the paradigm. We also introduce the seminal Bayes' rule that serves as a key component in Bayesian statistics. Then, building on this, we present the main ingredients of Bayesian inference in Section 1.2: the posterior distribution, the likelihood function, and the prior distribution. Section 1.3 provides some examples of simple cases where the prior distribution is chosen for algebraic convenience, giving rise to a closed-form expression for the posterior; these are called conjugate families in the literature. Finally, the last section of this chapter, Section 1.4, is dedicated to cases where we cannot get closed-form expressions and for which numerical integration is needed. Specifically, we discuss two influential Markov chain Monte Carlo samplers: the Gibbs sampler and the Metropolis–Hastings algorithm. We also discuss how to interpret the chains obtained by these methods (i.e., Markov chain diagnostics).

1.1 A Gentle Introduction to Bayesian Statistics

In Section 1.1, you learn how to:

- Describe qualitatively Bayesianism as an alternative to the frequentist approach.
- Give the historical context for Bayesian statistics.
- Use Bayes' rule to find conditional probabilities.
- Understand the basics of Bayesian statistics.

1.1.1 Bayesian versus Frequentist Statistics

Classic frequentist statistics rely on frequentist probability—an interpretation of probability in which an event's probability is defined as the limit of its relative frequency (or propensity) in many, repeatable trials. It draws conclusion from a sample that is one of many hypothetical datasets that could have been collected; the uncertainty is therefore due to the sampling error associated with the sample, while model parameters and various quantities of interest are fixed (but unknown to the experimenter).

Example 8.1.1. Coin Toss. Considering the simple case of coin tossing, if we flip a fair coin many times, we expect to see heads about 50% of the time. If we flip the coin only a few times, however, we could see a different distribution just by chance. Indeed, there is a non-zero probability of observing all heads if the sample is small enough. Figure 1.1 illustrates this very fact by showing the number of heads observed in 100 samples of five iid tosses; in this specific example, we observe six samples for which all tosses are heads.[1]

Yet, as the sample size increases, the relative frequency of heads should get closer to 50% if the coin is fair. Figure 1.2 reports that, if the number of tosses increases, then relative frequency of heads gets closer to 0.5—the probability of seeing heads on a given coin toss. In other words, increasing the sample size makes the sample less uncertain, and the experimenter should be reaching a probability of 0.5 in the limit, assuming they can reproduce the experiment an infinite number of times.

Bayesians see things differently: they interpret probabilities as degrees of certainty about some quantity of interest. To find such probabilities, they draw on prior knowledge about those quantities, expressing one's beliefs before some data are taken into account. Then, as data are collected, knowledge about the

[1] Each coin toss can be seen as a Bernoulli random variable, meaning that their sum is a binomial with parameters $q = 0.5$ and $m = 5$. See Chapter 19.1 for more details.

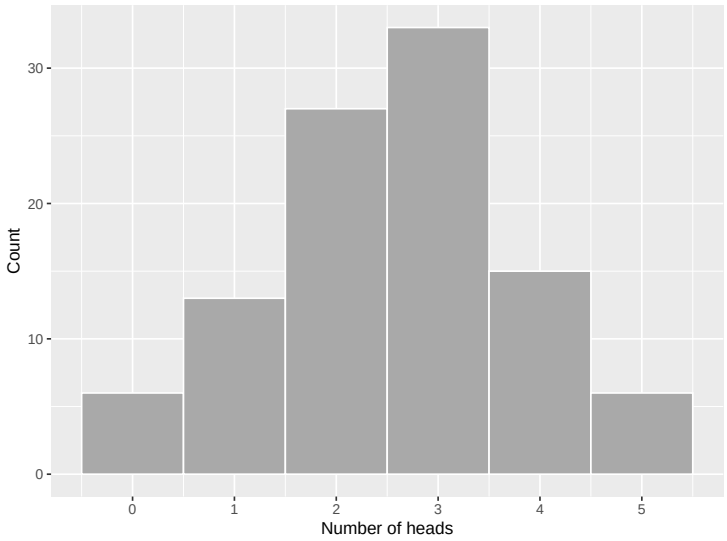

Figure 1.1: **Frequency histogram of the number of heads in a sample of five data points**

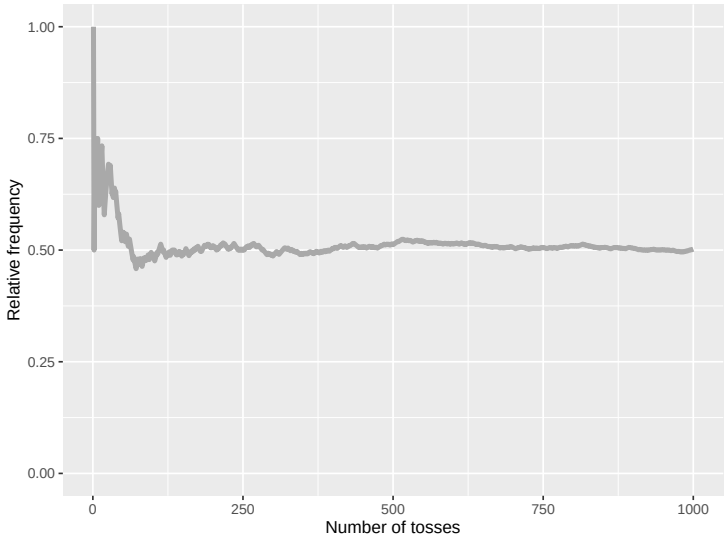

Figure 1.2: **Cumulative relative frequencies of heads for an increasing sample size**

world is updated, allowing us to incorporate such new information is a consistent manner; the resulting quantity is referred to as the posterior, which summarizes the information in both the prior and the data.

In the context of Bayesian inference and modeling, this interpretation of probability implies that model parameters are assumed to be random variables—unlike the frequentist approach that considers them fixed. Starting from the prior distribution, the data—summarized via the likelihood function—are used to update the prior distribution and create a posterior distribution of the parameters (see Section 1.2 for more details on the posterior distribution, the likelihood function, and the prior distribution). The influence of the prior distribution on the posterior distribution becomes weaker as the size of the observed data sample increases: the prior information is less and less relevant as new information comes in.

Example 8.1.1. Coin Toss, continued. We now reconsider the coin tossing experiment above through a Bayesian lens. Let us first assume that we have a (potentially unfair) coin, and we wish to understand the probability of obtaining heads, denoted by q in this example. Consistent with the Bayesian paradigm, this parameter is random; let us assume that the random variable associated with the probability of observing heads denoted by Q. For simplicity, we assume that we do not have prior information on the specific coin under investigation.[2] Assuming again that our sample contains only five iid tosses, we know that the probability of observing x heads is given by the binomial distribution with $m = 5$ such that

$$p_{X|Q=q}(x) = \Pr(X = x \,|\, Q = q) = \binom{5}{x} q^x (1-q)^{5-x}, \quad x \in \{0, 1, ..., 5\},$$

where $0 \leq q \leq 1$, which emphasizes the fact that this probability depends on parameter q by explicitly conditioning on it (unlike the notation used so far in this book, note that we append subscripts to the various pdf and pmf in this chapter to denote the random variables under study; this additional notation allows us to consider pdf and pmf of different random variables in the same problem).

Let us generate a sample of these five tosses:

```
set.seed(1)
nbheads <- c(1)
num_flips <- 5
coin <- c("heads", "tails")
flips <- sample(coin, size = 5, replace = TRUE)
```

[2] Specifically, we use a uniform over $[0, 1]$ for our prior distribution. As explained in Section 1.2.3, this type of prior is said to be noninformative.

```
nbheads <- sum(flips == "heads")
cat("Number of heads:", nbheads)
```

```
Number of heads: 3
```

Based on this simulation, we obtain a data sample that contains three heads and two tails. Therefore, using Bayesian statistics, we can show that

$$f_{Q|X=3}(q) \propto q^3(1-q)^2,$$

where $\propto$ means proportional to (note that obtaining this equation requires some tools that will be introduced in Section 1.2).[3] Figure 1.3 illustrates this pdf and reports the uncertainty about parameter q based on this sample of five data points. In this example, one can see that the uncertainty is quite large; this is a by-product of using only five data points. Indeed, based on these five observations, one could argue that the probability should be close to $\frac{3}{5} = 0.6$. This Bayesian analysis shows that 0.6 is likely, but that it is also very uncertain—a conclusion that is not direct in the frequentist approach.

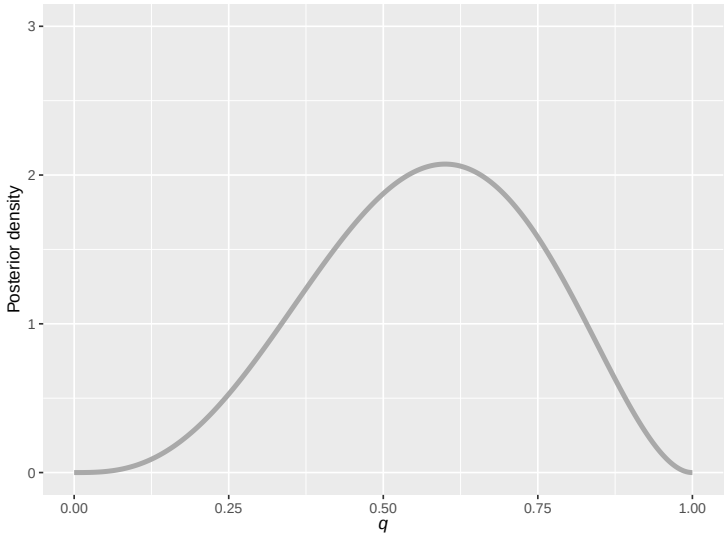

Figure 1.3: **Posterior probability density function of the parameter q for a sample of five data points**

Figure **??** reports the analog of Figure 1.2 through a Bayesian lens: we see the evolution of the posterior density of parameter q as a function of the sample size for the same sample used in Figure 1.2. As we obtain more evidence, the

[3] This is also an application of the beta–binomial conjugate family that will be explained in Section 1.3.1

posterior density becomes more concentrated around 0.5—a consequence of using a fair coin in the simulations above. Yet, even if the sample size if 1,000, we still see some parameter uncertainty.

But why be Bayesian? There are indeed several advantages to the Bayesian approach. First, this approach allows us to describe the entire distribution of parameters conditional on the data. This allows us, for example, to provide probability statements regarding the parameters that could be interpreted as such. Second, it provides a unified approach for estimating parameters. Some non-Bayesian methods, such as least squares, require a separate approach to estimate variance components. In contrast, in Bayesian methods, all parameters can be treated in a similar fashion. Third, it allows experimenters to blend prior information from other sources with the data in a coherent manner.[4]

1.1.2 A Brief History Lesson

Interestingly, some have argued that the birth of Bayesian statistics is intimately related to insurance; see, for instance, Cowles (2013). Specifically, the Great Fire of London in 1666—destroying more than 10,000 homes and about 100 churches—led to the rise of insurance as we know it today. Shortly after, the first full-fledged fire insurance company came into existence in England during the 1680s. By the turn of the century, the idea of insurance was well ingrained and its use was booming in England. Yet, the lack of statistical models and methods—much needed to understand risk—drove some insurers to bankruptcy.

Thomas Bayes, an English statistician, philosopher and Presbyterian minister, applied his mind to some of these important statistical questions raised by insurers. This culminated into Bayes' theory of probability in his seminar essay entitled *Essay towards solving a problem in the doctrine of chances*, published posthumously in 1763. This essay laid out the foundation of what we now know as Bayesian statistics.

Thomas Bayes' work also helped Pierre-Simon Laplace, a famous French scholar and polymath, to develop and popularize the Bayesian interpretation of probability in the late 1700s and early 1800s. He also moved beyond Bayes' essay and generalized his framework. Laplace's efforts were followed by many, and Bayesian thinking continued to progress throughout the years with the help of statisticians like Bruno de Finetti, Harold Jeffreys, Dennis Lindley, and Leonard Jimmie Savage.

Nowadays, Bayesian statistics and modeling is widely used in science, thanks to the increase in computational power over the past 30 years. Actuarial science and loss modeling, more specifically, have also been breeding grounds for

[4]There is also a rich history blending prior information with data in loss modeling and in actuarial science, generally speaking; it is known as credibility. For more details on experience rating using credibility theory, see Chapter 11.

Figure 1.4: **Portrait of an unknown Presbyterian clergyman identified as Thomas Bayes in O'Donnell (1936)**

Bayesian methodology. So, Bayesian statistics circles back to insurance, in a sense, where it all started.

1.1.3 Bayes' Rule

This subsection introduces how the Bayes' rule is applied to calculating conditional probability for events.

Conditional Probability. The concept of conditional probability considers the relationship between probabilities of two (or more) events happening. In its most simple form, being interested in conditional probability boils down to answering this question: *given that event B happened, how does this affect the probability that A happens?* To answer this question, we can define formally the concept of conditional probability:

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

To be properly defined, we must assume that $\Pr(B)$ is larger than zero; that is, event B is not impossible. Simply put, a conditional probability turns B into the new probability space, and then cares only about the part of A that is inside B (i.e., $A \cap B$).

Example 8.1.2. Actuarial Exam Question. An insurance company estimates that 40% of policyholders who have an extended health policy will renew next year, and 70% of policyholders who have a long-term disability policy will renew next year. The company also estimates that 50% of their clients who have both policies will renew at least one next year. The company records report that 65% of clients have an extended health policy, 40% have a long-term disability policy, and 10% have both. Using the data above, calculate the percentage of policyholders that will review at least one policy next year.

Solution. Let E be the event that a policyholder has an extended health policy, D be the event that a policyholder has a long-term disability policy, and R be the event that a policyholder renews a policy. We are given:

- $\Pr(E) = 0.65$,
- $\Pr(D) = 0.40$,
- $\Pr(E \cap D) = 0.10$,
- $\Pr(R \,|\, E \cap D^{\mathrm{c}}) = 0.40$,
- $\Pr(R \,|\, E^{\mathrm{c}} \cap D) = 0.70$,
- $\Pr(R \,|\, E \cap D) = 0.50$.

We are looking for $\Pr(R)$.

Note that

$$\Pr(E \cap D^{\mathrm{c}}) = \Pr(E) - \Pr(E \cap D) = 0.65 - 0.10 = 0.55,$$

and

$$\Pr(E^{\mathrm{c}} \cap D) = \Pr(D) - \Pr(E \cap D) = 0.40 - 0.10 = 0.30.$$

Moreover, note that $E \cap D^{\mathrm{c}}$, $E^{\mathrm{c}} \cap D$, and $E \cap D$ are mutually disjoint, and that

$$\begin{aligned}
\Pr(R) &= \Pr(R \cap (E \cap D^{\mathrm{c}})) + \Pr(R \cap (E^{\mathrm{c}} \cap D)) + \Pr(R \cap (E \cap D)) && (1.1)\\
&= \Pr(R \,|\, (E \cap D^{\mathrm{c}}))\Pr(E \cap D^{\mathrm{c}}) + \Pr(R \,|\, (E^{\mathrm{c}} \cap D))\Pr(E^{\mathrm{c}} \cap D) && (1.2)\\
&\quad + \Pr(R \,|\, (E \cap D))\Pr(E \cap D) && (1.3)\\
&= 0.40 \times 0.55 + 0.70 \times 0.30 + 0.50 \times 0.10 && (1.4)\\
&= 0.48. && (1.5)
\end{aligned}$$

Independence. If two events are unrelated to one another, we say that they are independent. Specifically, A and B are independent if

$$\Pr(A \cap B) = \Pr(A)\ \Pr(B).$$

For positive probability events, independence between A and B is also equivalent to

$$\Pr(A \,|\, B) = \Pr(A) \quad \text{and} \quad \Pr(B \,|\, A) = \Pr(B),$$

which means that the occurrence of event B does not have an impact on the occurence of A, and vice versa.

Bayes' Rule. Intuitively speaking, Bayes' rule provides a mechanism to put our Bayesian thinking into practice. It allows us to update our information by combining the data—from the likelihood—and the prior together to obtain a posterior probability.

Proposition 8.1.1. Bayes' Rule for Events. For events A and B, the posterior probability of event A given B follows

$$\Pr(A \mid B) = \frac{\Pr(B \mid A) \Pr(A)}{\Pr(B)},$$

where the law of total probability allows us to find

$$\Pr(B) = \Pr(A) \Pr(B \mid A) + \Pr(A^{\mathrm{c}}) \Pr(B \mid A^{\mathrm{c}}).$$

Note, again, that this works as long as event B is possible (i.e., $\Pr(B) > 0$).[5]

Proof. Bayes' rule may be derived from the definition of conditional probability shown above:

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

if $\Pr(B) > 0$. Similarly,

$$\Pr(B \mid A) = \frac{\Pr(A \cap B)}{\Pr(A)}$$

if $\Pr(A) > 0$. Solving for $\Pr(A \cap B)$ in the last equation and substituting into the first one yields Bayes' rule:

$$\Pr(A \mid B) = \frac{\Pr(B \mid A) \Pr(A)}{\Pr(B)}.$$

Simply put, the posterior probability of event A given B is obtained by combining the likelihood of B given a fixed A—proxied by $\Pr(B \mid A)$—with the prior probability of observing A, and then dividing it by the marginal probability of event B to make sure that the probabilities sum up to one.

Example 8.1.3. Actuarial Exam Question. An automobile insurance company insures drivers of all ages. An actuary compiled the following statistics on the company's insured drivers:

[5] The law of total probability states that the total probability of an event B is equal to the sum of the probabilities of B occurring under different conditions, weighted by the probabilities of those conditions. In the case where there are only two different conditions (let us say A and A^{c}), we simply need to consider these two conditions. In all generality, however, we would need to consider more possibilities if the sample space cannot be divided into only two events.

Age of Driver	Probability of Accident	Portion of Company's Insured Driver
16-20	0.06	0.08
21-30	0.03	0.15
31-65	0.02	0.49
66-99	0.04	0.28

A randomly selected driver that the company insures has an accident. Calculate the probability that the driver was age 16-20.

Solution. Let B be the event of an insured driver having an accident, and let

- A_1 be the event related to the driver's age being in the range 16-20,
- A_2 be the event related to the driver's age being in the range 21-30,
- A_3 be the event related to the driver's age being in the range 31-65,
- A_4 be the event related to the driver's age being in the range 66-99.

Then,

$$\Pr(A_1 \mid B) = \frac{\Pr(B \mid A_1)\Pr(A_1)}{\Pr(B \mid A_1)\Pr(A_1) + \Pr(B \mid A_2)\Pr(A_2) + \Pr(B \mid A_3)\Pr(A_3) + \Pr(B \mid A_4)\Pr(A_4)} \tag{1.6}$$

$$= \frac{0.06 \times 0.08}{0.06 \times 0.08 + 0.03 \times 0.15 + 0.02 \times 0.49 + 0.04 \times 0.28} \tag{1.7}$$

$$= 0.1584. \tag{1.8}$$

1.1.4 An Introductory Example of Bayes' Rule

The example above illustrates how to use Bayes' rule in an academic context; the focus of this book is, nonetheless, data analytics. We therefore also wish to illustrate Bayes' rule by using *real* data. In this introductory example, we use the Singapore auto data `sgautonb` of the R package `CASdatasets` that was already used in Chapter **??** (see also Section **??** for more details).

```
library(CASdatasets)
data(sgautonb)
```

This dataset contains information about the number of car accidents and some risk factors (i.e., the type of the vehicle insured, the age of the vehicle, the sex of the policyholder, and the age of the policyholder grouped into seven categories).[6]

[6]The data are from the General Insurance Association of Singapore, an organization consisting of non-life insurers in Singapore. These data contains the number of car accidents for $n = 7{,}483$ auto insurance policies with several categorical explanatory variables and the exposure for each policy.

Example 8.1.4. Singapore Insurance Data. A new insurance company—targeting an older segment of the population—estimates that 20% of their policyholders will be 65 years old and older. The actuaries working at the insurance company believes that the Singapore insurance dataset is credible to understand the accident occurrence of the new company. Based on this information, find the probability that a randomly selected driver having (at least) an accident is 65 years old and older for the new insurance company.

Solution. Let O denote the event related to the policyholder being 65 years old and older (i.e., Age Category 6 in the dataset), and A the event of a policyholder having at least an accident. Using Bayes' rule, we have that

$$\Pr(O \,|\, A) = \frac{\Pr(A \,|\, O)\Pr(O)}{\Pr(A)},$$

where the prior probability $\Pr(O)$ is given by the problem statement: $\Pr(O) = 0.20$. This implies that $\Pr(O^c) = 1 - 0.20 = 0.80$. From the Singapore insurance data, we know that $\Pr(A \,|\, O) = 0.1082803$ and $\Pr(A \,|\, O^c) = 0.06415506$, which allow us to use the law of total probability to obtain:

$$\Pr(A) = \Pr(A \,|\, O)\Pr(O) + \Pr(A \,|\, O^c)\Pr(O^c).$$

```
n <- length(sgautonb$AgeCat)
nO <- sum(sgautonb$AgeCat == 6)
nOc <- sum(sgautonb$AgeCat != 6)
nAandO <- sum(sgautonb$AgeCat == 6 & sgautonb$Clm_Count > 0)
nAandOc <- sum(sgautonb$AgeCat != 6 & sgautonb$Clm_Count > 0)

PAO <- nAandO/nO
PAOc <- nAandOc/nOc

POA <- PAO * 0.2/(PAO * 0.2 + PAOc * 0.8)
cat("The probability that policyholder having accident is 65 years old and older is",
    POA)
```

```
The probability that policyholder having accident is 65 years old and older is 0.2967391
```

The probability that a randomly selected driver has (at least) an accident is 65 years old and older is therefore about 29.7% for the new insurance company. Simply put, we started with an *a priori* probability of 20%, meaning that unconditionally, we should have about 20% of policyholders aged 65 years old and older, and that 20% of the policyholders should have at least one accident. Then, based on the observed data, this probability is updated to 29.7%: the data seem to imply that, of all people having accidents, there are more older policyholders than what we would have guessed just based on our prior assumption.

In the next section, we will expand on the idea of Bayes' rule and apply it to slightly more general cases involving random variables instead of events.

1.2 Building Blocks of Bayesian Inference

In Section 1.2, you learn how to:

- Describe the main components of Bayesian inference; that is, the posterior distribution, the likelihood function, and the prior distribution.
- Summarize the different classes of prior used in practice.

Proposition 8.1.1 above deals with the elementary case of Bayes' rule for events. Although this version of Bayes' rule is useful to understand the foundation of Bayesian statistics, we will need slightly more general versions of it to achieve Bayesian inference. Specifically, Proposition 8.1.1 needs to be generalized to the case of random variables.

Let us first consider the case of discrete random variables. Assume X and Y are both discrete random variables that allow for the following joint pmf of

$$p_{X,Y}(x, y) = \Pr(X = x \text{ and } Y = y)$$

as well as the following marginal distributions for X and Y:

$$p_X(x) = \Pr(X = x) = \sum_k p_{X,Y}(x, k) \quad \text{and} \quad p_Y(y) = \Pr(Y = y) = \sum_k p_{X,Y}(k, y),$$

respectively. Using the result of Proposition 8.1.1 and setting event A as $\{Y = y\}$ and B as $\{X = x\}$ yields

$$p_{Y|X=x}(y) = \frac{p_{X|Y=y}(x)\, p_Y(y)}{p_X(x)},$$

where $p_{Y|X=x}(y) = \Pr(Y = y \,|\, X = x)$ is the conditional pmf of Y conditional on X being equal to x. Using the law of total probability,

$$p_X(x) = \sum_k p_{X,Y}(x, k) = \sum_k p_{X|Y=k}(x)\, p_Y(k),$$

we can rewrite the denominator above to get the following version of Bayes' rule:

$$p_{Y|X=x}(y) = \frac{p_{X|Y=y}(x)\, p_Y(y)}{\sum_k p_{X|Y=k}(x)\, p_Y(k)}.$$

We can also obtain a similar Bayes' rule for continuous random variables by replacing probability mass functions by probability density functions, and sums by integrals.

Proposition 8.2.1. Bayes' Rule for Continuous Random Variables. For two continuous random variables X and Y, the conditional probability density function of Y given $X = x$ follows

$$f_{Y|X=x}(y) = \frac{f_{X|Y=y}(x)\, f_Y(y)}{f_X(x)},$$

where the marginal distributions of X and Y are given as follows:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,u)\, du \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(u,y)\, du,$$

respectively. Similar to the discrete random variable case, we can swap the denominator of the equation above for

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,u)\, du = \int_{-\infty}^{\infty} f_{X|Y=u}(x)\, f_Y(u)\, du$$

by using the law of total probability.

Proof. Bayes' rule for continuous random variables may be derived from the definition of conditional probability density functions:

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)},$$

if $f_X(x) > 0$. Similarly,

$$f_{X|Y=y}(x) = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

if $f_Y(y) > 0$. Solving for $f_{X,Y}(x,y)$ in the last equation and substituting into the first one yields Bayes' rule for continuous random variables:

$$f_{Y|X=x}(y) = \frac{f_{X|Y=y}(x)\, f_Y(y)}{f_X(x)}.$$

Note that one can mix the discrete and continuous definitions of Bayes' rule to accommodate for cases where the parameters have continuous random variables and the observations are expressed via discrete random variables, or vice versa.

1.2.1 Posterior Distribution

Model parameters are assumed to be random variables under the Bayesian paradigm, meaning that Bayes' rule for (discrete or continuous) random variables can be applied to update the prior knowledge about parameters by using new data. This is indeed similar to the process used in Section 1.1.1.

Let us consider only one (random) model parameter θ associated with random variable Θ for now.[7] Further, consider n observations

$$\mathbf{x} = (x_1, x_2, ..., x_n),$$

which are realizations of the collection of random variables

$$\mathbf{X} = (X_1, X_2, ..., X_n).$$

If Y in Proposition 8.2.1 is replaced by Θ and X by $\mathbf{X}$, we obtain

$$f_{\Theta|\mathbf{X}=\mathbf{x}}(\theta) = \frac{f_{\mathbf{X}|\Theta=\theta}(\mathbf{x})\, f_{\Theta}(\theta)}{f_{\mathbf{X}}(\mathbf{x})},$$

which represents the posterior distribution of the model parameter after updating the distribution based on the new observations $\mathbf{x}$, and where

- $f_{\mathbf{X}|\Theta=\theta}(\mathbf{x})$ is the likelihood function, also known as the conditional joint pdf of the observations assuming a given value of parameter θ,
- $f_{\Theta}(\theta)$ is the unconditional pdf of the parameter that represents the prior information, and
- $f_{\mathbf{X}}(\mathbf{x})$ is the marginal likelihood, which is a constant term with respect to θ, making the posterior density integrate to one.

In other words, when applied to Bayesian inference, Bayes' rule provides a mean to update the prior distribution of the parameter into a posterior distribution—by considering the observations $\mathbf{x}$.

Note that the marginal likelihood is constant once we have the observations. It does not depend on θ and does not impact the overall shape of the pdf: it only provides the adequate scaling to ensure that the density integrates to one.

[7] For the sake of simplicity, we only consider one parameter in our derivation here. Note that, later, we will consider cases with more than one parameter and that this extension does not change the bulk of our results and derivations.

For this reason, it is common to write down the posterior distribution using a proportional relationship instead:

$$f_{\Theta|\mathbf{X}=\mathbf{x}}(\theta) \propto \underbrace{f_{\mathbf{X}|\Theta=\theta}(\mathbf{x})}_{\text{Likelihood}} \; \underbrace{f_{\Theta}(\theta)}_{\text{Prior}}.$$

Example 8.2.1. A Problem Inspired from Meyers (1994). A car insurance pays the following (independent) claim amounts on an automobile insurance policy:

$$1050, \qquad 1250, \qquad 1550, \qquad 2600, \qquad 5350, \qquad 10200.$$

The amount of a single payment is distributed as a single-parameter Pareto distribution with $\theta = 1000$ and α unknown, such that

$$f_{X_i|A=\alpha}(x_i) = \frac{\alpha\, 1000^{\alpha}}{x_i^{\alpha+1}}, \quad x_i \in \mathbb{R}_+.$$

We assume that the prior distribution of α is given by a gamma distribution with shape parameter 2 and scale parameter 1, and its pdf is given by

$$f_A(\alpha) = \alpha\, e^{-\alpha}, \quad \alpha \in \mathbb{R}_+.$$

Find the posterior distribution of parameter α.

Solution. The likelihood function is constructed by multiplying the pdf of the single payment amounts because they are independent; that is,

$$f_{\mathbf{X}|A=\alpha}(\mathbf{x}) = \prod_{i=1}^{6} f_{X_i|A=\alpha}(x) = \frac{\alpha^6\, 1000^{6\alpha}}{\prod_{i=1}^{6} x_i^{\alpha+1}} = \alpha^6\, e^{-5.66518\alpha - 41.44653}.$$

The posterior distribution is given by

$$f_{A|\mathbf{X}=\mathbf{x}}(\alpha) = \frac{\alpha^7\, e^{-6.66518\alpha - 41.44653}}{\int_0^{\infty} \alpha^7\, e^{-6.66518\alpha - 41.44653}\, d\alpha} = \frac{\alpha^7\, e^{-6.66518\alpha}}{\int_0^{\infty} \alpha^7\, e^{-6.66518\alpha}\, d\alpha}.$$

Interestingly, we do not need to solve the integral at the denominator to find this distribution. As we know that the results should be a proper pdf and that the numerator looks like a gamma distribution, we can deduce that

$$f_{A|\mathbf{X}=\mathbf{x}}(\alpha) = \frac{6.66518^8}{\Gamma(8)}\, \alpha^7\, e^{-6.66518\,\alpha},$$

which is a gamma distribution with shape parameter 8 and scale parameter $\frac{1}{6.66518}$. Figure 1.5 reports the posterior distribution of α.

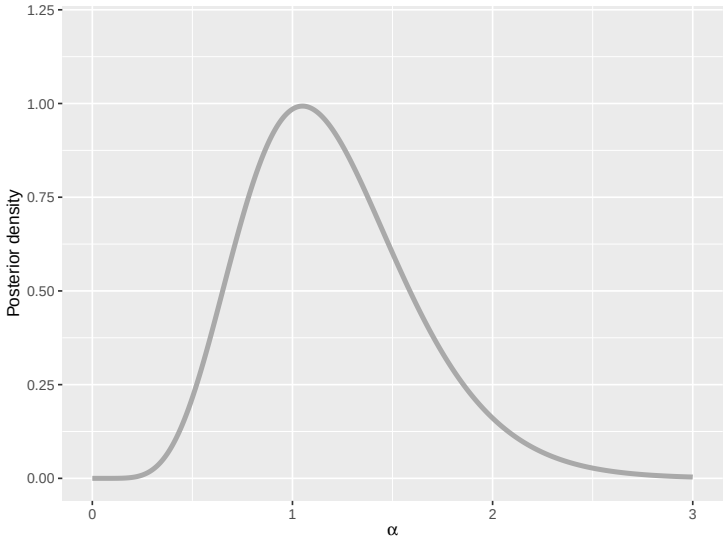

Figure 1.5: **Posterior densities of parameter** α

The discussion above considered continuous random variables, but the same logic can be applied to discrete random variables by replacing probability density functions by probability mass functions.

Example 8.2.2. Coin Toss Revisited. Assume that you observe three heads out of five (independent) tosses. Each toss has a probability of q of observing heads and $1-q$ of observing tails. Find the posterior distribution of q assuming a uniform prior distribution over the interval $[0, 1]$.

Solution. The prior distribution of q is given by

$$f_Q(q) = 1, \quad q \in [0, 1].$$

Assuming the likelihood function conditional on $Q = q$ is given by a binomial distribution with $m = 5$ and $x = 3$,

$$p_{X|Q=q}(x) = \binom{5}{3} q^3 (1-q)^2,$$

we have that the posterior distribution of q is given by

$$f_{Q|X=3}(q) \propto p_{X|Q=q}(x)\, f_Q(q) = q^3\,(1-q)^2,$$

which is a beta distribution with $a = 4$, $b = 3$, and $\theta = 1$; that is, we can easily deduce that

$$f_{Q|X=3}(q) = \frac{\Gamma(7)}{\Gamma(4)\Gamma(3)}\, q^3\,(1-q)^2.$$

In the following subsections, we will discuss at greater length the two main building blocks used to build the posterior distribution: the likelihood function and the prior distribution.

1.2.2 Likelihood Function

The likelihood function is a fundamental concept in statistical inference. It is used to estimate the parameters of a statistical model based on observed data. As mentioned in previous chapters, the likelihood function can be used to find the maximum likelihood estimator. In Bayesian statistics, the likelihood function is used to update the prior based on the evidence (or data).

As explained above and in Appendix C, the likelihood function is defined as the conditional joint pdf or pmf of the observed data, given the model parameters. In other words, it is the probability of observing the data given a specific parameter values.

Mathematically, the likelihood function is written as $f_{\mathbf{X}|\Theta=\theta}(x)$ (for continuous random variables) or $p_{\mathbf{X}|\Theta=\theta}(x)$ (for discrete random variables). Note that, throughout the book, the notation $L(\theta|\mathbf{x})$ has also been used for the likelihood function, and we will use both interchangeably in this chapter.

Special Case: Independent and Identically Distributed Observations. Oftentimes, in many problems and real-world applications, the observations are assumed to be iid. If they are, then we can easily write the likelihood function as:

$$f_{\mathbf{X}|\Theta=\theta}(\mathbf{x}) = \prod_{i=1}^{n} f_{X_i|\Theta=\theta}(x_i) \quad \text{or} \quad p_{\mathbf{X}|\Theta=\theta}(\mathbf{x}) = \prod_{i=1}^{n} p_{X_i|\Theta=\theta}(x_i).$$

1.2.3 Prior Distribution

In the Bayesian paradigm, the prior distribution represents our knowledge or beliefs about the unknown parameters before we observe any data. It is a probability distribution that expresses the uncertainty about the values of the parameters. The prior distribution is typically specified by choosing a family of probability distributions and selecting specific values for its parameters.

The choice of prior distribution is subjective and often based on external information or previous studies. In some cases, noninformative priors can be used, which represent minimal prior knowledge or assumptions about the parameters. In other cases, informative and weakly informative priors can be used, which incorporate prior knowledge or assumptions based on external sources. The selection of the prior distribution should be carefully considered, and sensitivity analysis can be performed to assess the robustness of the results to different prior assumptions.

Why Does It Matter? The choice of prior distribution can have a significant impact on the results of a Bayesian analysis. Different prior distributions can lead to different posterior distributions, which are the updated probability distributions for the parameters after we observe the data. Therefore, it is important to choose a prior distribution that reflects our prior knowledge or beliefs about the parameters.

Informative and Weakly Informative Priors

Informative and weakly informative priors are terms used to describe the amount of prior knowledge or beliefs that is incorporated into a statistical model. Informative priors contain substantial prior knowledge about the parameters of a model, while weakly informative priors contain moderate prior knowledge.

Informative priors are useful when there is strong, potentially subjective prior information available about the model parameters, which can help to constrain the posterior distribution and improve inference. For example, in an insurance claims analysis study, an informative prior may be used to incorporate previous knowledge, such as the results of a previous claims study. This can help to reduce the uncertainty in the estimation of the claim distribution and improve the power of the analysis.

On the other hand, weakly informative priors are used when there is some—yet little—prior knowledge available or when the goal is to allow the data to drive the inference. Weakly informative priors are designed to mildly impact the posterior distribution and are often chosen based on principles such as symmetry or scale invariance.

Overall, the choice of prior depends on the specific problem at hand and the available prior knowledge or beliefs. Informative priors can be useful when prior information is available and can improve the precision of the posterior distribution. In contrast, weakly informative priors can be useful when the goal is to allow the data to drive the inference and avoid imposing strong prior assumptions.

Example 8.2.3. Actuarial Exam Question. You are given:

- Annual claim frequencies follow a Poisson distribution with mean λ.

- The prior distribution of λ has the following pdf:

$$f_\Lambda(\lambda) = (0.3)\frac{1}{6}e^{-\frac{\lambda}{6}} + (0.7)\frac{1}{12}e^{-\frac{\lambda}{12}}, \quad \text{where } \lambda > 0.$$

Ten claims are observed for an insured in Year 1. Calculate the expected value of the posterior distribution of λ.

Solution. The posterior distribution can be found from:

$$\begin{aligned} f_{\Lambda|X=10}(\lambda) &= \frac{p_{X|\Lambda=\lambda}(10) f_\Lambda(\lambda)}{p_X(10)} && (1.9)\\ &= \frac{\frac{e^{-\lambda}\lambda^{10}}{10!}\left((0.5)\frac{1}{6}e^{-\frac{\lambda}{6}} + (0.5)\frac{1}{12}e^{-\frac{\lambda}{12}}\right)}{\int_0^\infty \frac{e^{-\lambda}\lambda^{10}}{10!}\left((0.5)\frac{1}{6}e^{-\frac{\lambda}{6}} + (0.5)\frac{1}{12}e^{-\frac{\lambda}{12}}\right) d\lambda} && (1.10)\\ &= \frac{\lambda^{10}\left(\frac{0.5}{6}e^{-\frac{7\lambda}{6}} + \frac{0.5}{12}e^{-\frac{13\lambda}{12}}\right)}{118170}. && (1.11) \end{aligned}$$

The posterior mean is therefore given by

$$\begin{aligned} \mathrm{E}\left[\Lambda \mid X = 10\right] &= \frac{1}{118170}\int_0^\infty \lambda^{11}\left(\frac{0.5}{6}e^{-\frac{7\lambda}{6}} + \frac{0.5}{12}e^{-\frac{13\lambda}{12}}\right) d\lambda && (1.12)\\ &= \frac{1}{118170}\left(\frac{0.5}{6}(11!)(6/7)^{12} + \frac{0.5}{12}(11!)(12/13)^{12}\right) && (1.13)\\ &= 9.81328. && (1.14) \end{aligned}$$

Noninformative Priors

It is possible to take the idea of weakly informative priors to the extreme by using noninformative priors. A noninformative prior is a prior distribution that is intentionally chosen to allow the data to have a more decisive influence on the posterior distribution rather than being overly influenced by prior beliefs or assumptions.

Noninformative priors can take different forms, such as flat priors, for instance. A flat prior assigns equal probability to all possible parameter values without additional information or assumptions.

Example 8.2.4. Informative Versus Noninformative Priors. You wish to investigate the impact of having informative and noninformative priors on a claim frequency analysis. Assume that the claim frequency for each policy follows a Bernoulli random variable with a probability of q such that

$$q_{X_i|Q=q}(x_i) = q^{x_i}(1-q)^{1-x_i}, \quad x_i \in \{0,1\},$$

where $q \in [0,1]$, and consider two different prior distributions:

- Informative: Based on past experience, you know that the claim probability is typically less than 5%, thus justifying the use of a uniform distribution over $[0, 0.05]$.
- Noninformative: You do not wish your posterior distribution to be impacted by your prior assumption and simply select a uniform distribution over the domain of p, which is $[0,1]$.

Using the first 250 lines of the Singapore insurance dataset (see Example 8.1.4 for more details on this dataset), find the two posterior distributions as well as the posterior expected value of the probability q under both prior assumptions.

Solution. Let us start with the informative prior, where

$$f_Q(q) = \frac{1}{0.05 - 0} = 20, \quad \text{if } q \in [0, 0.05],$$

and zero otherwise. In this case, assuming $x = \sum_{i=1}^{250} x_i$, the posterior density is given by

$$\begin{aligned} f_{Q|\mathbf{X}=\mathbf{x}}(q) &\propto f_{\mathbf{X}|Q=q}(\mathbf{x}) f_Q(q) && (1.15)\\ &\propto \prod_{i=1}^{250} q^{x_i}(1-q)^{1-x_i} && (1.16)\\ &= q^x(1-q)^{250-x)}, \quad \text{if } 0 \le q \le 0.05, && (1.17) \end{aligned}$$

and zero otherwise. We can numerically obtain the shape of this posterior distribution by dividing $q^x(1-q)^{250-x}$ by

$$\int_0^{0.05} q^x(1-q)^{250-x}\, dq.$$

The second prior is still uniform, but over $[0,1]$ this time, which is given mathematically by

$$f_Q(q) = \frac{1}{1-0} = 1, \quad \text{if } q \in [0,1],$$

and zero otherwise, leading to the following posterior distribution:

$$f_{Q|\mathbf{X}=\mathbf{x}}(q) \propto f_{\mathbf{X}|Q=q}(\mathbf{x}) f_Q(q) \tag{1.18}$$

$$\propto \prod_{i=1}^{250} q^{x_i}(1-q)^{1-x_i} \tag{1.19}$$

$$= q^x(1-q)^{250-x}, \quad \text{if } 0 \leq q \leq 1, \tag{1.20}$$

and zero otherwise.

```
qs <- seq(from = 0, to = 0.12, by = 1e-04)
x <- sum(sgautonb$Clm_Count[1:250])

integrandposterior1 <- function(q) {
    q^x * (1 - q)^(250 - x) * ifelse(q >= 0 & q <= 0.05, 1, 0)
}
marglikelihood1 <- integrate(integrandposterior1, 0, 1, abs.tol = .Machine$double.eps^2)$value
posterior1 <- integrandposterior1(qs)/marglikelihood1

integrandposterior2 <- function(q) {
    q^x * (1 - q)^(250 - x) * ifelse(q >= 0 & q <= 1, 1, 0)
}
marglikelihood2 <- integrate(integrandposterior2, 0, 1, abs.tol = .Machine$double.eps^2)$value
posterior2 <- integrandposterior2(qs)/marglikelihood2
```

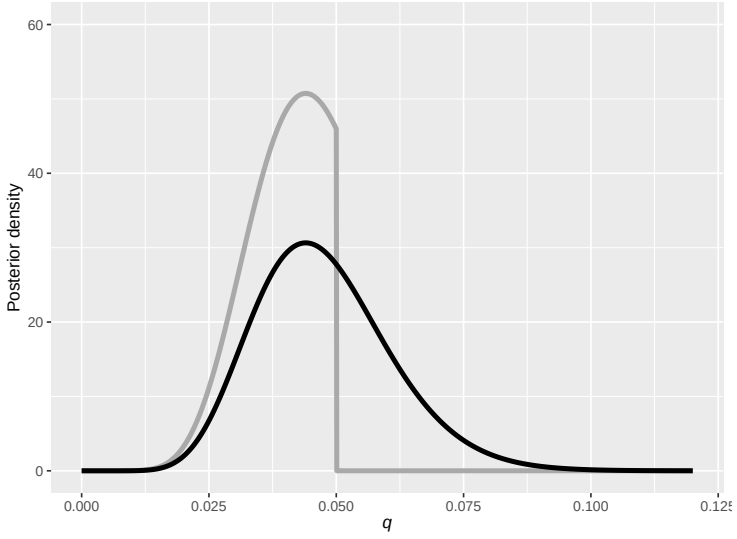

Figure 1.6: **Posterior densities based on informative (gray) and noninformative priors (black)**

We also wish to obtain the expected value of q for both posterior distribution. This can be obtained by numerically integrating the following equation:

$$\mathrm{E}[Q|\mathbf{X}=\mathbf{x}] = \int_0^1 q\, f_{Q|\mathbf{X}=\mathbf{x}}(q)\, dq.$$

```
integrandexpvalue1 <- function(q) {
    integrandposterior1(q)/marglikelihood1 * q
}
expectedvalue1 <- integrate(integrandexpvalue1, 0, 1, abs.tol = .Machine$double.eps^2)$
cat("The posterior expected value of the parameter when using the informative prior is"
    expectedvalue1)
```

```
The posterior expected value of the parameter when using the informative prior is 0.038
```

```
integrandexpvalue2 <- function(q) {
    integrandposterior2(q)/marglikelihood2 * q
}
expectedvalue2 <- integrate(integrandexpvalue2, 0, 1, abs.tol = .Machine$double.eps^2)$
cat("The posterior expected value of the parameter when using the noninformative prior
    expectedvalue2)
```

```
The posterior expected value of the parameter when using the noninformative prior is 0
```

As one can see, these values are different, meaning that the prior distribution can have a material impact on the posterior distribution. One should therefore be careful when selecting a prior distribution.

Improper Priors

An improper prior is a prior distribution that is not a proper probability distribution, meaning that it does not integrate (or sum) to one over the entire parameter space. Improper priors can be used in Bayesian analyses, but they require careful handling because they can lead to improper posterior distributions.

Improper priors are typically used when there is little or no prior information about the parameter of interest—some noninformative priors are indeed improper—and they can be thought of as representing a very diffuse or noncommittal prior belief. For instance, the uniform distribution on an infinite interval is a common choice of improper prior.

Example 8.2.5. Improper Prior, Proper Posterior. Let us assume a random sample $\mathbf{x}$ of size n, which is a realization of the collection of random

variables $\mathbf{X} = (X_1, X_2, ..., X_n)$. Further, assume that each random variable X_i is independent and normally distributed with mean of μ and variance of 1:

$$f_{X_i|M=\mu}(x_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \mu)^2\right), \quad x_i \in \mathbb{R},$$

where μ is a (random) parameter. Obtain the posterior distribution of μ assuming that its prior distribution is improper and given by $f_M(\mu) \propto 1$, where $\mu \in \mathbb{R}$.

Solution. According to Bayes' rule, we have that

$$f_{M|\mathbf{X}=\mathbf{x}}(\mu) = \frac{f_{\mathbf{X}|M=\mu}(\mathbf{x})\, f_M(\mu)}{f_{\mathbf{X}}(\mathbf{x})} \propto \prod_{i=1}^{n} f_{X_i|M=\mu}(x_i)$$

because $f_M(\mu) \propto 1$ and $f_{\mathbf{X}}(\mathbf{x})$ does not depend on μ. Using the equation above, we can obtain the posterior distribution by simplifying the following equation:

$$\begin{aligned}
f_{M|\mathbf{X}=\mathbf{x}}(\mu) &\propto \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right) && (1.21)\\
&\propto \exp\left(-\frac{1}{2}\left(\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2\right)\right) && (1.22)\\
&\propto \exp\left(-\frac{n}{2}\left(\frac{\sum_{i=1}^{n} x_i^2}{n} - \frac{2\mu\sum_{i=1}^{n} x_i}{n} + \mu^2\right)\right) && (1.23)\\
&\propto \exp\left(-\frac{n}{2}\left(-\frac{2\mu\sum_{i=1}^{n} x_i}{n} + \mu^2\right)\right) && (1.24)\\
&\propto \exp\left(-\frac{n}{2}\left(\mu - \frac{\sum_{i=1}^{n} x_i}{n}\right)^2\right) && (1.25)\\
&\propto \frac{1}{\sqrt{2\pi\frac{1}{n}}} \exp\left(-\frac{1}{2}\frac{\left(\mu - \frac{\sum_{i=1}^{n} x_i}{n}\right)^2}{\frac{1}{n}}\right), && (1.26)
\end{aligned}$$

which is a normal distribution with mean $\frac{\sum_{i=1}^{n} x_i}{n}$ and variance $\frac{1}{n}$. Interestingly, this posterior distribution is proper even though the prior distribution was improper.

1.3 Conjugate Families

In Section 1.3, you learn how to:

- Describe three specific classes of conjugate families.
- Use conjugate distributions to determine posterior distributions of parameters.
- Understand the pros and cons of conjugate family models.

In Bayesian statistics, if a posterior distribution is the same distribution as the prior distribution, the prior and posterior are called conjugate distributions. Note that both posterior and prior have similar shapes but will have different parameters, generally speaking.

But Why? Two main reasons explain why conjugate families have been so popular historically:

1. They are easy to use from a computational standpoint: posterior distributions in most conjugate families can be obtained in closed form, making this class of models easy to use even if we do not have access to computing power.
2. They tend to be easy to interpret: posterior distributions are compromises between data and prior distributions. Having both prior and posterior distributions in the same family—but with different parameters—allows us to understand and quantify how the data changed our initial assumptions.

1.3.1 The Beta–Binomial Conjugate Family

The first conjugate family that we investigate in this book is the beta–binomial family. Let $\mathbf{X} = (X_1, X_2, ...X_m)$ represent a sample of iid Bernoulli random variables such that

$$X_i = \begin{cases} 1 & \text{if success} \\ 0 & \text{if failure} \end{cases},$$

with probabilities q and $1 - q$, respectively. Let us further define $x = \sum_{i=1}^{m} x_i$ the sum of the realized successes.

We know from elementary probability that $X = \sum_{i=1}^{m} X_i$ follows a binomial distribution (i.e., the number of successes x in m Bernoulli trials) with unknown probability of success q in $[0, 1]$, similar to the coin tossing case of Example 8.1.1, such that the likelihood function is given by

$$p_{\mathbf{X}|Q=q}(\mathbf{x}) = \binom{m}{x} q^x (1-q)^{m-x}, \quad x \in \{0, 1, ..., m\},$$

where $x = \sum_{i=1}^{m} x_i$. The latter represents our evidence. Then, we combine it with its usual conjugate prior—the beta distribution with parameters a and b. The pdf of the beta distribution is given as follows:

$$f_Q(q) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} q^{a-1}(1-q)^{b-1}, \quad q \in [0,1],$$

where a and b are shape parameters of the beta distribution.[8]

We can now combine the prior distribution—beta—with the likelihood function—binomial—to obtain the posterior distribution.

Proposition 8.3.1. Beta–Binomial Conjugate Family. Consider a sample of m iid Bernoulli experiments $(X_1, X_2, ..., X_m)$ each with success probability q. Further assume that the random variable associated with the success probability, Q, has a prior that is beta with shape parameters a and b. The posterior distribution of Q is therefore given by

$$f_{Q|\mathbf{X}=\mathbf{x}}(q) = \frac{\Gamma(a+b+m)}{\Gamma(a+x)\Gamma(b+m-x)} q^{a+x-1}(1-q)^{b+m-x-1},$$

where $x = \sum_{i=1}^{m} x_i$, which is a beta distribution with shape parameters $a+x$ and $b+m-x$.

Proof. From Section 1.2.1, we know that

$$\begin{aligned} f_{Q|\mathbf{X}=\mathbf{x}}(q) = \frac{p_{\mathbf{X}|Q=q}(\mathbf{x})\, f_Q(q)}{p_{\mathbf{X}}(\mathbf{x})} &\propto \binom{m}{x} q^x (1-q)^{m-x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} q^{a-1}(1-q)^{b-1} \\ &\propto q^{a+x-1}(1-q)^{b+m-x-1}. \end{aligned}$$

We therefore only need to find the normalizing constant that ensures that the right-hand of the equation above is a density. Interestingly, the right-hand side looks like a beta distribution; specifically,

[8] Here, we assume that the domain of the beta is $[0,1]$, meaning that $\theta = 1$. For more details, see Appendix D.

$$\int_0^1 q^{a+x-1}(1-q)^{b+m-x-1}\,dq \tag{1.27}$$

$$= \frac{\Gamma(a+x)\Gamma(b+m-x)}{\Gamma(a+b+m)} \int_0^1 \frac{\Gamma(a+b+m)}{\Gamma(a+x)\Gamma(b+m-x)} q^{a+x-1}(1-q)^{b+m-x-1}\,dq \tag{1.28}$$

$$= \frac{\Gamma(a+x)\Gamma(b+m-x)}{\Gamma(a+b+m)}, \tag{1.29}$$

and

$$f_{Q|\mathbf{X}=\mathbf{x}}(q) = \frac{\Gamma(a+b+m)}{\Gamma(a+x)\Gamma(b+m-x)} q^{a+x-1}(1-q)^{b+m-x-1}.$$

Parameters Versus Hyperparameters. In this context, a and b are called hyperparameters—parameters of the prior. These are different from parameters of the underlying model (i.e., q in the beta–binomial family). Hyperparameters are typically known and determined by the experimenter, whereas the underlying model parameters are random in the Bayesian context.

Example 8.3.1. Actuarial Exam Question. You are given:

- The annual number of claims in Year i for a policyholder has a binomial distribution with pmf

$$p_{X_i|Q=q}(x_i) = \binom{2}{x} q^{x_i}(1-q)^{2-x_i}, \quad x_i \in \{0,1,2\}.$$

- The prior distribution is

$$f_Q(q) = 4q^3, \quad q \in [0,1].$$

The policyholder had one claim in each of Years 1 and 2. Calculate the Bayesian estimate of the expected number of claims in Year 3.

Solution. The likelihood function based on this policyholder's number of claims in Years 1 and 2 is given by:

$$p_{\mathbf{X}|Q=q}(\mathbf{x}) = p_{X_1|Q=q}(1)\, p_{X_2|Q=q}(1) = \binom{2}{1} q^1(1-q)^1 \binom{2}{1} q^1(1-q)^1 \propto q^2(1-q)^2,$$

which is proportional to a binomial pmf with $m = 4$, two successes, and a success probability of q. Because the prior distribution is beta distributed with $a = 4$ and $b = 1$, we know that the posterior distribution of parameter q is given by

$$\begin{aligned} f_{Q|\mathbf{X}=\mathbf{x}}(q) &= \frac{\Gamma(4+1+4)}{\Gamma(4+2)\Gamma(1+4-2)} q^{4+2-1}(1-q)^{1+4-2-1} \\ &= \frac{\Gamma(9)}{\Gamma(6)\Gamma(3)} q^5(1-q)^2 \\ &= 168q^5(1-q)^2, \end{aligned}$$

which is also a beta distribution with shape parameters 6 and 3, respectively.

The expected number of claim in Year 3 is

$$\mathrm{E}\left[\mathrm{E}\left[X_3 \,|\, Q = q\,\right]\right] = \mathrm{E}\left[2q\right] = 2\,\mathrm{E}\left[q\right],$$

and $\mathrm{E}[q]$ is the expected value of the beta distribution, which is given by

$$\mathrm{E}\left[q\right] = \frac{6}{6+3} = \frac{2}{3}.$$

Ultimately, this leads to an expected number of claim in Year 3 of $2\left(\frac{2}{3}\right) = \frac{4}{3}$.

Example 8.3.2. Impact of Beta Prior on Posterior. You wish to investigate the impact of having different beta hyperparameters on the posterior distribution. Assume that the claim frequency for each policy follows a Bernoulli random variable with a probability of q such that

$$p_{X_i|Q=q}(x_i) = q^{x_i}(1-q)^{1-x_i}, \quad x_i \in \{0,1\},$$

where $q \in [0,1]$, and consider two different sets of hyperparameters:

- Set 1: $a = 1$ and $b = 10$.
- Set 2: $a = 2$ and $b = 2$.

Figure 1.7 shows the pdf of these two prior distributions.

Using again the first 250 lines of the Singapore insurance dataset (see Example 8.1.4 for more details on this dataset), find the two posterior distributions.

Solution. The likelihood function associated with the observations is given by

$$p_{\mathbf{X}|Q=q}(\mathbf{x}) = \binom{250}{x} q^x(1-q)^{250-x}, \quad \text{where } x = \sum_{i=1}^{250} x_i,$$

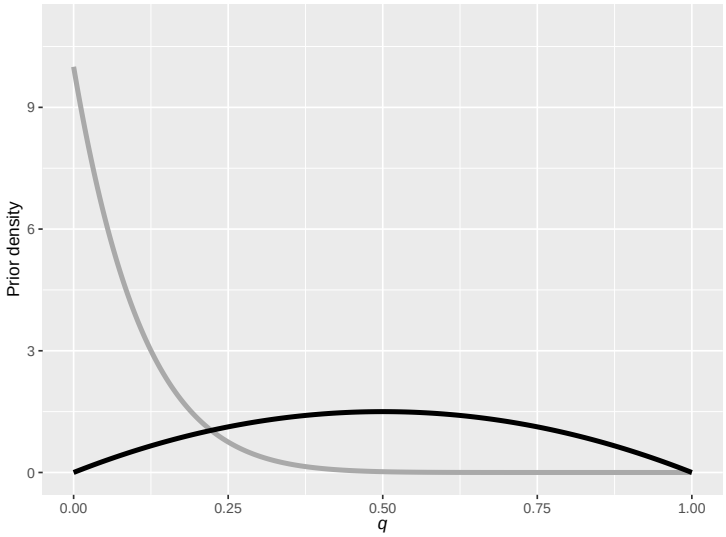

Figure 1.7: **Beta prior densities:** $a = 1$ **and** $b = 10$ **(gray), and** $a = 2$ **and** $b = 2$ **(black)**

as mentioned already in Example 8.2.4. Combining this likelihood with a beta prior gives a beta posterior:

$$f_{Q|\mathbf{X}=\mathbf{x}}(q) = \frac{\Gamma(a+b+250)}{\Gamma(a+x)\,\Gamma(b+250-x)}\, q^{a+x-1}(1-q)^{b+250-x-1},$$

that can be evaluated for various values of a and b. Figure 1.8 reports the two posterior distributions associated with the priors mentioned above.

```
x <- sum(sgautonb$Clm_Count[1:250])

posterior1 <- dbeta(qs, shape1 = 1 + x, shape2 = 10 + 250 - x)
posterior2 <- dbeta(qs, shape1 = 2 + x, shape2 = 2 + 250 - x)

dataposterior <- data.frame(x = qs, y1 = posterior1, y2 = posterior2)

ggplot(dataposterior, aes(x = x, y = y1)) + geom_line(color = "darkgray", lwd = 1.5) +
    geom_line(aes(y = y2), color = "black", lwd = 1.5) + xlim(0, 1) + ylim(0, 35) +
    xlab(expression(italic("q"))) + ylab("Posterior density")
```

The prior distribution (and its hyperparameters) clearly has an impact on the posterior distribution. As a general rule of thumb for the beta prior, a higher

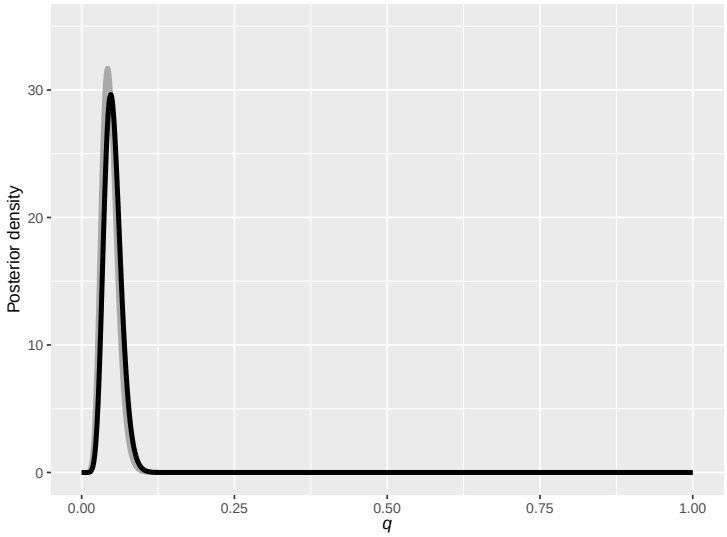

Figure 1.8: **Posterior densities based on two different priors: $a = 1$ and $b = 10$ (gray), and $a = 2$ and $b = 2$ (black)**

a puts more weight on higher values of q and a higher b puts more weight on lower values of q.

1.3.2 The Gamma–Poisson Conjugate Family

We now present a second conjugate family: the gamma–Poisson family. Let $\mathbf{X} = (X_1, X_2, ..., X_n)$ be a sample of iid Poisson random variables such that

$$p_{X_i|\Lambda=\lambda}(x_i) = \frac{\lambda^{x_i}\, e^{-\lambda}}{x_i!}, \quad x_i \in \mathbb{R}_+.$$

The likelihood function associated with this sample would therefore be given by

$$f_{\mathbf{X}|\Lambda=\lambda}(\mathbf{x}) = \prod_{i=1}^{n} p_{X_i|\Lambda=\lambda}(x_i) = \prod_{i=1}^{n} \frac{\lambda^{x_i}\, e^{-\lambda}}{x_i!} = \frac{\lambda^{x}\, e^{-n\lambda}}{\prod_{i=1}^{n} x_i!} \propto \lambda^{x}\, e^{-n\lambda},$$

where $x = \sum_{i=1}^{n} x_i$. The shape of this likelihood function, as a function of λ, is reminiscent of a gamma distribution, hinting to the fact that this distribution would be a good contender for a conjugate prior. Indeed, if we let the prior distribution be gamma with shape hyperparameter α and scale hyperparameter θ,

$$f_\Lambda(\lambda) = \frac{1}{\Gamma(\alpha)\theta^\alpha}\lambda^{\alpha-1}\, e^{-\frac{\lambda}{\theta}}, \quad \lambda \in \mathbb{R}_+,$$

we can show that the posterior distribution of λ is also gamma.

Proposition 8.3.2. Gamma–Poisson Conjugate Family. Consider a sample of n iid Poisson experiments $(X_1, X_2, ..., X_n)$, each with rate parameter λ. Further assume that the random variable associated with the rate, Λ, has a prior that is gamma distributed with shape hyperparameter α and scale hyperparameter θ. The posterior distribution of Λ is therefore given by

$$f_{\Lambda|\mathbf{X}=\mathbf{x}}(\lambda) = \frac{1}{\Gamma(\alpha+x)\left(\frac{\theta}{n\theta+1}\right)^{\alpha+1}}\lambda^{\alpha+x-1}\, e^{-\frac{\lambda(n\theta+1)}{\theta}},$$

where $x = \sum_{i=1}^n x_i$, which is a gamma distribution with shape parameter $\alpha + x$ and scale parameter $\frac{\theta}{n\theta+1}$.

Proof. From Section 1.2.1, we know that

$$\begin{aligned} f_{\Lambda|\mathbf{X}=\mathbf{x}}(\lambda) = \frac{p_{\mathbf{X}|\Lambda=\lambda}(\mathbf{x})\, f_\Lambda(\lambda)}{p_{\mathbf{X}}(\mathbf{x})} &\propto \lambda^x\, e^{-n\lambda}\, \lambda^{\alpha-1}\, e^{-\frac{\lambda}{\theta}} \\ &\propto \lambda^{\alpha+x-1}\, e^{-\frac{\lambda(n\theta+1)}{\theta}}, \end{aligned}$$

where $x = \sum_{i=1}^n x_i$. We therefore only need to find the normalizing constant that ensures that the right-hand of the equation above is a density. Interestingly, the right-hand side looks like a gamma distribution; specifically,

$$\int_0^\infty \lambda^{\alpha+x-1}\, e^{-\frac{\lambda(n\theta+1)}{\theta}}\, d\lambda \tag{1.30}$$

$$= \Gamma(\alpha+x)\left(\frac{\theta}{n\theta+1}\right)^{\alpha+1} \int_0^\infty \frac{1}{\Gamma(\alpha+x)\left(\frac{\theta}{n\theta+1}\right)^{\alpha+1}}\, \lambda^{\alpha+x-1}\, e^{-\frac{\lambda(n\theta+1)}{\theta}}\, d\lambda \tag{1.31}$$

$$= \Gamma(\alpha+x)\left(\frac{\theta}{n\theta+1}\right)^{\alpha+1}, \tag{1.32}$$

and

$$f_{\Lambda|\mathbf{X}=\mathbf{x}}(\lambda) = \frac{1}{\Gamma(\alpha+x)\left(\frac{\theta}{n\theta+1}\right)^{\alpha+1}}\lambda^{\alpha+x-1}\, e^{-\frac{\lambda(n\theta+1)}{\theta}}.$$

Example 8.3.3. Actuarial Exam Question. You are given:

- The number of claims incurred in a month by any insured has a Poisson distribution with mean λ.
- The claim frequencies of different insured are iid.
- The prior distribution is gamma with pdf

$$f_\Lambda(\lambda) = \frac{(100\lambda)^6}{120\lambda} e^{-100\lambda}, \quad \lambda \in \mathbb{R}_+.$$

- The number of claims every month is distributed as follows:

Month	Number of Insured	Number of Claims
1	100	6
2	150	8
3	200	11
4	300	?

Calculate the expected number of claims in Month 4.

Solution. The likelihood function based on this policyholder's number of claims in Months 1, 2, and 3 is given by:

$$p_{\mathbf{X}|\Lambda=\lambda}(\mathbf{x}) = p_{X_1|\Lambda=\lambda}(6)\, p_{X_2|\Lambda=\lambda}(8)\, p_{X_3|\Lambda=\lambda}(11) \propto \lambda^{6+8+11}\, e^{-\lambda(100+150+200)}.$$

Because the prior distribution is gamma distributed with $\alpha = 6$ and $\theta = \frac{1}{100}$, we know that the posterior distribution of parameter λ is also gamma distributed with shape parameter

$$\alpha + x = 6 + 6 + 8 + 11 = 31$$

and scale parameter

$$\frac{\theta}{n\theta + 1} = \frac{\frac{1}{100}}{(100 + 150 + 200)\frac{1}{100} + 1} = \frac{1}{550}.$$

The expected number of claim in Month 4 is

$$\mathrm{E}\left[\mathrm{E}\left[X_4 \,|\, \Lambda = \lambda\right]\right] = \mathrm{E}\left[300\lambda\right] = 300\,\mathrm{E}\left[\lambda\right],$$

and $\mathrm{E}\left[\lambda\right]$ is the expected value of the gamma distribution, which is given by

$$\mathrm{E}[\lambda] = \frac{31}{550}.$$

Ultimately, this leads to an expected number of claim in Month 4 of $300\left(\frac{31}{550}\right) = \frac{930}{55} \approx 16.91$.

1.3.3 The Normal–Normal Conjugate Family

The last conjugate family is the normal–normal family. Let $\mathbf{X} = (X_1, X_2, ..., X_n)$ be a sample of iid normal random variables such that

$$f_{X_i|M=\mu}(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right), \quad x_i \in \mathbb{R}.$$

Further, to keep our focus on μ, we will assume throughout our analysis that the variance parameter σ^2 is known.[9] The likelihood function associated with this sample would therefore be given by

$$f_{\mathbf{X}|M=\mu}(\mathbf{x}) = \prod_{i=1}^{n} f_{X_i|M=\mu}(x_i) \tag{1.33}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}\right) \tag{1.34}$$

$$\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}\right). \tag{1.35}$$

A very natural prior distribution that matches the likelihood structure is unsurprisingly the normal distribution. Let us assume that the prior distribution for μ is given by

$$f_M(\mu) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2}\frac{(\mu-\theta)^2}{\tau^2}\right),$$

where θ is the mean parameter and τ^2 is the variance parameter. We can then easily show that the posterior distribution of μ is also given by a normal distribution.

---

[9] Conjugate families for the normal distribution with unknown σ^2 can also be derived. For the sake of simplicity, we will only focus on the case with known variance parameter in this book.

Proposition 8.3.3. Normal–Normal Conjugate Family. Consider a sample of n iid normals $(X_1, X_2, ..., X_n)$, each with mean parameter μ and variance parameter σ^2 that is known. Further assume that the random variable associated with the mean, M, has a prior that is normally distributed with mean hyperparameter θ and variance hyperparameter τ^2. The posterior distribution of M is therefore given by

$$f_{M|\mathbf{X}=\mathbf{x}}(\mu) = \frac{1}{\sqrt{2\pi\left(\frac{\tau^2\sigma^2}{n\tau^2+\sigma^2}\right)}} \exp\left(-\frac{1}{2}\frac{\left(\mu - \left(\frac{x}{n}\frac{\tau^2}{n\tau^2+\sigma^2} + \theta\frac{\sigma^2}{n\tau^2+\sigma^2}\right)\right)^2}{\frac{\tau^2\sigma^2}{n\tau^2+\sigma^2}}\right),$$

where $x = \sum_{i=1}^{n} x_i$, which is a normal distribution with mean parameter

$$\frac{x}{n}\frac{n\tau^2}{n\tau^2+\sigma^2} + \theta\frac{\sigma^2}{n\tau^2+\sigma^2}$$

and variance parameter

$$\frac{\tau^2\sigma^2}{n\tau^2+\sigma^2}.$$

Proof. From Section 1.2.1, we know that

$$\begin{aligned}
f_{M|\mathbf{X}=\mathbf{x}}(\mu) &= \frac{f_{\mathbf{X}|M=\mu}(\mathbf{x})\, f_M(\mu)}{f_{\mathbf{X}}(\mathbf{x})} \qquad (1.36)\\
&\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}\right)\exp\left(-\frac{1}{2}\frac{(\mu-\theta)^2}{\tau^2}\right)\\
&\propto \exp\left(-\frac{1}{2}\frac{\sum_{i=1}^{n}x_i^2 - 2\mu x + n\mu^2}{\sigma^2} - \frac{1}{2}\frac{\mu^2 - 2\mu\theta + \theta^2}{\tau^2}\right)\\
&\propto \exp\left(-\frac{1}{2}\frac{n\mu^2 - 2\mu x}{\sigma^2} - \frac{1}{2}\frac{\mu^2 - 2\mu\theta}{\tau^2}\right)\\
&\propto \exp\left(-\frac{1}{2}\frac{\mu^2\left(n\tau^2+\sigma^2\right) - 2\mu\tau^2 x - 2\mu\sigma^2\theta}{\tau^2\sigma^2}\right)\\
&\propto \exp\left(-\frac{1}{2}\frac{\mu^2 - 2\mu\left(x\frac{\tau^2}{n\tau^2+\sigma^2} + \theta\frac{\sigma^2}{n\tau^2+\sigma^2}\right)}{\frac{\tau^2\sigma^2}{n\tau^2+\sigma^2}}\right)\\
&\propto \frac{1}{\sqrt{2\pi\left(\frac{\tau^2\sigma^2}{n\tau^2+\sigma^2}\right)}} \exp\left(-\frac{1}{2}\frac{\left(\mu - \left(\frac{x}{n}\frac{n\tau^2}{n\tau^2+\sigma^2} + \theta\frac{\sigma^2}{n\tau^2+\sigma^2}\right)\right)^2}{\frac{\tau^2\sigma^2}{n\tau^2+\sigma^2}}\right),
\end{aligned}$$

where $x = \sum_{i=1}^{n} x_i$.

The prior distribution hyperparameters and posterior distribution parameters can be interpreted in the normal–normal conjugate family:

- For the prior, θ represents the *a priori* value of the mean parameter, and τ^2 is related to the precision of that prior mean (i.e., the larger the value, the less precise the prior mean is, and vice versa).
- For the posterior, the new mean parameter is a weighted average between the prior mean parameter θ and the sample mean $\frac{x}{n}$. The new variance parameter is informed by the prior variability τ^2 and the variability of the data σ^2.

Example 8.3.4. Impact of Normal Prior on Posterior. Assume the following observed automobile claims for a small portfolio of policies:

$$1050, \qquad 1250, \qquad 1550, \qquad 2600, \qquad 5350, \qquad 10200.$$

Further assume that the logarithm of the claim amount follows a normal distribution with parameters μ and $\sigma^2 = 1$. Find the posterior distribution of the mean parameter μ for a normal prior distribution where $\theta = 7$. Consider different values of τ^2; that is, $\tau^2 = 0.1$, $\tau^2 = 1$, and $\tau^2 = 10$. Figure 1.9 shows the pdf of these three prior distributions.

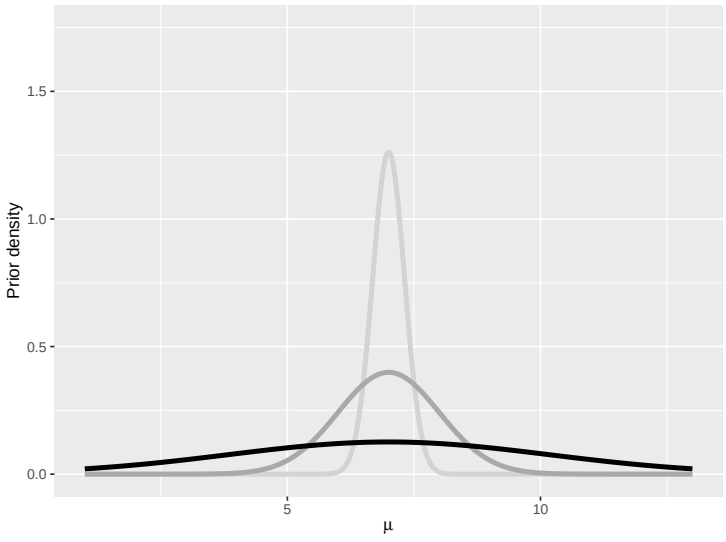

Figure 1.9: **Normal prior densities: $\tau^2 = 0.1$ (light gray), $\tau^2 = 1$ (gray), and $\tau^2 = 10$ (black)**

Solution. Using the results of Proposition 8.3.3, we can obtain the following posterior distributions:

```
xi <- c(1050, 1250, 1550, 2600, 5350, 10200)
x <- sum(log(xi))
n <- length(xi)
sigma2 <- 1

mean1 <- theta * (sigma2/(n * tau21 + sigma2)) + x/n * ((n * tau21)/(n * tau21 +
    sigma2))
mean2 <- theta * (sigma2/(n * tau22 + sigma2)) + x/n * ((n * tau22)/(n * tau22 +
    sigma2))
mean3 <- theta * (sigma2/(n * tau23 + sigma2)) + x/n * ((n * tau23)/(n * tau23 +
    sigma2))

var1 <- (tau21 * sigma2)/(n * tau21 + sigma2)
var2 <- (tau22 * sigma2)/(n * tau22 + sigma2)
var3 <- (tau23 * sigma2)/(n * tau23 + sigma2)

posterior1 <- dnorm(xs, mean = mean1, sd = sqrt(var1))
posterior2 <- dnorm(xs, mean = mean2, sd = sqrt(var2))
posterior3 <- dnorm(xs, mean = mean3, sd = sqrt(var3))

dataposterior <- data.frame(x = xs, y1 = posterior1, y2 = posterior2, y3 = posterior3)

ggplot(dataposterior, aes(x = x, y = y1)) + geom_line(color = "lightgray", lwd = 1.5) +
    geom_line(aes(y = y2), color = "darkgray", lwd = 1.5) + geom_line(aes(y = y3),
    color = "black", lwd = 1.5) + xlim(1, 13) + ylim(0, 1.75) + xlab(expression(italic("q"))) +
    ylab("Posterior density")
```

Interestingly, as shown in Example 8.3.4, the prior distribution can have some impact on the final posterior distribution. When the prior assumption about the mean is very precise, having a few data points do not create a huge gap between the prior and the posterior (see the light gray curves in Figures 1.9 and 1.10). When the prior is very imprecise, on the other hand, then the data are allow to speak, and the posterior can be quite different from the prior distribution.

1.3.4 Criticism of Conjugate Family Models

While conjugate family models have some advantages, such as ease of interpretation and computational simplicity, they also have some limitations:

1. Conjugate families are oftentimes chosen for their mathematical convenience rather than their ability to accurately model the data under study. This can lead to models that are too simplistic and lack the flexibility

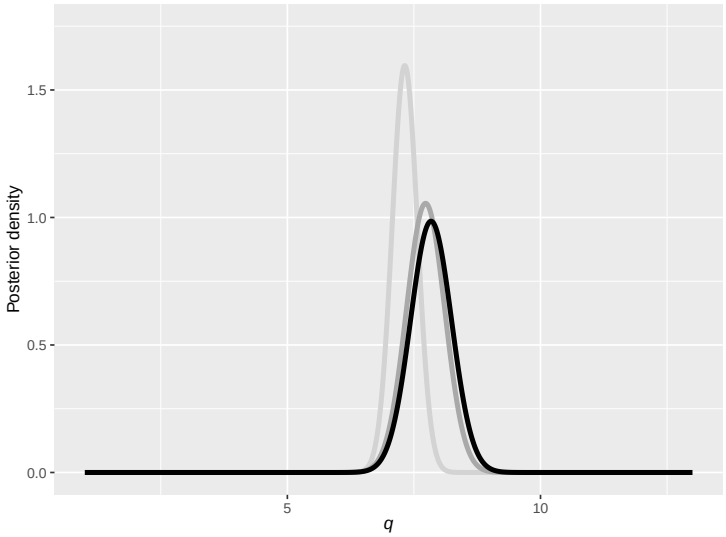

Figure 1.10: **Posterior densities based on three different priors: $\tau^2 = 0.1$ (light gray), $\tau^2 = 1$ (gray), and $\tau^2 = 10$ (black)**

needed to model real-world phenomena.
2. Conjugate family models rely on the choice of prior distribution, and different choices can lead to very different posterior distributions.
3. Conjugate family models are only applicable to a narrow range of problems, which limit their usefulness in practical applications.

It is important to note that while conjugate family models have their limitations, they can still be useful in certain situations, especially when the assumptions of the model are well understood and the data are relatively simple.

1.4 Posterior Simulation

In Section 1.4, you learn how to:

- Use the standard computational tools for Bayesian inference.
- Diagnose Markov chain convergence.

1.4.1 Introduction to Markov Chain Monte Carlo Methods

Sometimes, using conjugate family models is ill-suited for the problem at hand, and more complicated priors need to be selected. Under other circumstances, complex models involve many parameters making the posterior distribution intractable. In these cases, the posterior distribution of the parameters will not have a closed-form solution, generally speaking, and will need to be estimated via numerical methods.

A common way to generate draws of the parameter posterior distribution is to create Markov chains for which their stationary distributions correspond to the posterior of interest. These Markov chain-based methods are known as Markov chain Monte Carlo (MCMC) methods in the literature. This section provides a brief overview of these methods and of their uses. We do not intent to give much of the theory behind these methods, which would require a deep understanding of Markov chains and their theory.[10] Instead, we focus on their applications in insurance and loss modeling. Specifically, in the next two subsections, we introduce the two most common MCMC methods; that is, the Gibbs sampler of Gelfand and Smith (1990) and the Metropolis–Hastings algorithm of Hastings (1970) and Metropolis et al. (1953).

1.4.2 The Gibbs Sampler

As mentioned above, sometimes, we cannot use conjugate families. In other cases where the parameter space is large, it can be very hard to find the marginal likelihood $f_{\mathbf{X}}(\mathbf{x})$ (also known as the normalizing constant); that is, assuming that the model parameters are given by $\boldsymbol{\theta} = [\,\theta_1 \quad ...\theta_2 \quad ... \quad \theta_k\,]$ and contains k parameters, the marginal likelihood given by

$$f_{\mathbf{X}}(\mathbf{x}) = \int \int ... \int f_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}}(\mathbf{x}) f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})\, d\theta_1\, d\theta_2 \ldots d\theta_k$$

is hard to compute even when using typical quadrature-based rules, especially if k is large.

Fortunately, under very mild regulatory conditions, samples of the joint estimates of parameters can be obtained by sequentially sampling each parameter individually and by keeping all the other parameters constant. To do so, the distribution of any given parameter conditional on all the other parameters (and the data) needs to be known. These distributions are known as full conditional distributions; that is,

$$f_{\Theta_i \,|\, \mathbf{X}=\mathbf{x}, \boldsymbol{\Theta}_{\setminus i}=\boldsymbol{\theta}_{\setminus i}}(\theta_i),$$

[10] For an overview of the theory behind MCMC methods, see Robert and Casella (1999).

for parameter θ_i, where $\boldsymbol{\theta}_{\setminus i}$ represents all parameters except for the i^{th} one, and $\boldsymbol{\Theta}_{\setminus i}$ is the random variable associated with this set of parameters.

The full conditional distribution is an important building block in Gibbs sampling. Indeed, if one can obtain each parameter's distribution conditional on having the value of all the other parameters in closed form, then it is possible to generate samples for each parameter. Specifically, starting from an arbitrary set of starting values $\boldsymbol{\theta}^{(0)} = [\,\theta_1^{(0)} \quad \theta_2^{(0)} \quad ... \quad \theta_k^{(0)}\,]$, samples for each parameter can be generated by performing the following steps for $m = 1, 2, ..., M$:

1. Draw $\theta_1^{(m)}$ from $f_{\Theta_1 \,|\, \mathbf{X}=\mathbf{x}, \Theta_2=\theta_2^{(m-1)}, ..., \Theta_k=\theta_k^{(m-1)}}(\theta_1)$.
2. Draw $\theta_2^{(m)}$ from $f_{\Theta_2 \,|\, \mathbf{X}=\mathbf{x}, \Theta_1=\theta_1^{(m)}, \Theta_3=\theta_3^{(m-1)}, ..., \Theta_k=\theta_k^{(m-1)}}(\theta_2)$.
3. Draw $\theta_3^{(m)}$ from $f_{\Theta_3 \,|\, \mathbf{X}=\mathbf{x}, \Theta_1=\theta_1^{(m)}, \Theta_2=\theta_2^{(m)}, \Theta_4=\theta_4^{(m-1)}, ..., \Theta_k=\theta_k^{(m-1)}}(\theta_3)$.

   $\vdots$

k. Draw $\theta_k^{(m)}$ from $f_{\Theta_k \,|\, \mathbf{X}=\mathbf{x}, \Theta_1=\theta_1^{(m)}, ..., \Theta_{k-1}=\theta_{k-1}^{(m)}}(\theta_k)$.

The sample, especially at first, will depend on the initial values, $\boldsymbol{\theta}^{(0)}$, and it might take some time until the sampler can explore fully the distribution. For this reason, in practice, experimenters discard the first M^* iterations to make sure their analysis is not impacted by the choice of initial parameter; this initial period of discarded sample is known as the burn-in period.

The rest of the sample—the remaining $M - M^*$ iterations—is kept to estimate the posterior distribution and any quantities of interest.

Application to Bayesian Linear Regression

In statistics and in its most simple form, a linear regression is an approach for modeling the relationship between a scalar response and an explanatory variable. The former quantity is denoted by x_i for $i \in \{1, ..., n\}$, and the latter quantity is denoted by z_i for $i \in \{1, ..., n\}$ in this chapter. Mathematically, we can write this relationship as:

$$x_i = \alpha + \beta z_i + \varepsilon_i,$$

where ε_i is a disturbance term that captures the potential for errors in the linear relationship. This error term is typically assumed to be normally distributed with mean zero and variance σ^2.

In general, the coefficients α and β are unknown and need to be estimated. The experimenter can rely on Bayesian inference to find out the posterior distribution of the parameters α and β along with that of σ^2.

For the rest of the subsection, we investigate a specific application of Gibbs sampling to the context of linear regression.

We begin by computing the likelihood function conditional on the parameter values:

$$f_{\mathbf{X}\mid A=\alpha,\, B=\beta,\, \Sigma^2=\sigma^2}(\mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\alpha-\beta z_i)^2}{2\sigma^2}\right)$$
$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\alpha-\beta z_i)^2}{2\sigma^2}\right),$$

which is the first building block to construct our posterior distribution.

Then, we need a prior distribution, which could be informative, weakly informative, or noninformative. In this application, we select a prior that allows us to obtain each parameter's full conditional distribution in closed form. Specifically, we use a normal distribution for α and β, and an inverse gamma distribution for σ^2 with shape parameter $\frac{n_\sigma}{2}$ and scale parameter $\frac{\theta_\sigma}{2}$, where

$$f_A(\alpha) = \frac{1}{\sqrt{2\pi\tau_\alpha^2}} \exp\left(-\frac{1}{2}\frac{(\alpha-\theta_\alpha)^2}{\tau_\alpha^2}\right),$$
$$f_B(\beta) = \frac{1}{\sqrt{2\pi\tau_\beta^2}} \exp\left(-\frac{1}{2}\frac{(\beta-\theta_\beta)^2}{\tau_\beta^2}\right),$$
$$f_{\Sigma^2}(\sigma^2) = \frac{(\theta_\sigma/2)^{n_\sigma/2}}{\Gamma(n_\sigma/2)} \left(\frac{1}{\sigma^2}\right)^{n_\sigma/2+1} \exp\left(-\frac{\theta_\sigma/2}{\sigma^2}\right).$$

Proposition 8.4.1. Full Conditional Distributions of Bayesian Linear Regression Parameters. Consider a sample of n observations $\mathbf{x} = (x_1, ..., x_n)$ for which

$$x_i = \alpha + \beta z_i + \varepsilon_i,$$

where ε_i is normally distributed with mean zero and variance σ^2. The full conditional distributions of parameters α, β, and σ^2 are given by the following expressions:

$$A \sim \text{Normal}\left(\frac{1}{n}\left(\sum_{i=1}^{n} x_i - \beta z_i\right)\frac{n\tau_\alpha^2}{n\tau_\alpha^2+\sigma^2} + \theta_\alpha \frac{\sigma^2}{n\tau_\alpha^2+\sigma^2}, \frac{\tau_\alpha^2\sigma^2}{n\tau_\alpha^2+\sigma^2}\right),$$

$$B \sim \text{Normal}\left(\frac{1}{n}\left(\sum_{i=1}^{n} z_i\left(x_i-\alpha\right)\right)\frac{n\tau_\beta^2}{\tau_\beta^2\sum_{i=1}^{n} z_i^2+\sigma^2} + \theta_\beta \frac{\sigma^2}{\tau_\beta^2\sum_{i=1}^{n} z_i^2+\sigma^2}, \frac{\tau_\beta^2\sigma^2}{\tau_\beta^2\sum_{i=1}^{n} z_i^2+\sigma^2}\right),$$

$$\Sigma^2 \sim \text{Inverse Gamma}\left(\frac{n_\sigma+n}{2}, \frac{\theta_\sigma+\sum_{i=1}^{n}\left(y_i-\alpha-\beta z_i\right)^2}{2}\right),$$

respectively, assuming the prior distributions mentioned above.

Proof. From Section 1.2, we know that

$$f_{A,B,\Sigma^2|\mathbf{X}=\mathbf{x}}(\alpha,\beta,\sigma^2) \propto f_{\mathbf{X}|A=\alpha,\,B=\beta,\,\Sigma^2=\sigma^2}(\mathbf{x})\, f_A(\alpha)\, f_B(\beta)\, f_{\Sigma^2}(\sigma^2),$$

which is useful to derive the full conditional distributions of α, β, and σ^2.

Let us begin with α:

$$\begin{aligned}
&f_{A|\mathbf{X}=\mathbf{x},\,B=\beta,\,\Sigma^2=\sigma^2}(\alpha)\\
&\propto \exp\left(-\frac{1}{2}\frac{\sum_{i=1}^{n}\left(x_i-\alpha-\beta z_i\right)^2}{\sigma^2}\right)\exp\left(-\frac{1}{2}\frac{(\alpha-\theta_\alpha)^2}{\tau_\alpha^2}\right)\\
&\propto \exp\left(-\frac{1}{2}\left(\frac{n\alpha^2-2\alpha\sum_{i=1}^{n}(x_i-\beta z_i)}{\sigma^2}+\frac{\alpha^2-2\alpha\theta_\alpha}{\tau_\alpha^2}\right)\right)\\
&\propto \exp\left(-\frac{1}{2}\left(\frac{\alpha^2-2\alpha\left(\frac{1}{n}\left(\sum_{i=1}^{n}x_i-\beta z_i\right)\frac{n\tau_\alpha^2}{n\tau_\alpha^2+\sigma^2}+\theta_\alpha\frac{\sigma^2}{n\tau_\alpha^2+\sigma^2}\right)}{\frac{\tau_\alpha^2\sigma^2}{n\tau_\alpha^2+\sigma^2}}\right)\right)\\
&\propto \frac{1}{\sqrt{2\pi\left(\frac{\tau_\alpha^2\sigma^2}{n\tau_\alpha^2+\sigma^2}\right)}}\exp\left(-\frac{1}{2}\left(\frac{\left(\alpha-\left(\frac{1}{n}\left(\sum_{i=1}^{n}x_i-\beta z_i\right)\frac{n\tau_\alpha^2}{n\tau_\alpha^2+\sigma^2}+\theta_\alpha\frac{\sigma^2}{n\tau_\alpha^2+\sigma^2}\right)\right)^2}{\frac{\tau_\alpha^2\sigma^2}{n\tau_\alpha^2+\sigma^2}}\right)\right)
\end{aligned}$$

which is a normal distribution with mean parameter

$$\frac{1}{n}\left(\sum_{i=1}^{n}x_i-\beta z_i\right)\frac{n\tau_\alpha^2}{n\tau_\alpha^2+\sigma^2}+\theta_\alpha\frac{\sigma^2}{n\tau_\alpha^2+\sigma^2}$$

and variance parameter

$$\frac{\tau_\alpha^2\sigma^2}{n\tau_\alpha^2+\sigma^2}.$$

The derivation to obtain the full conditional distribution of β is similar to that of α:

$$
\begin{aligned}
& f_{B|\mathbf{X}=\mathbf{x},\, A=\alpha,\, \Sigma^2=\sigma^2}(\beta) \\
& \propto \exp\left(-\frac{1}{2}\frac{\sum_{i=1}^{n}\left(x_i-\alpha-\beta z_i\right)^2}{\sigma^2}\right)\exp\left(-\frac{1}{2}\frac{(\beta-\theta_\beta)^2}{\tau_\beta^2}\right) \\
& \propto \exp\left(-\frac{1}{2}\left(\frac{\beta^2\sum_{i=1}^{n}z_i^2-2\beta\sum_{i=1}^{n}z_i(x_i-\alpha)}{\sigma^2}+\frac{\beta^2-2\beta\theta_\beta}{\tau_\beta^2}\right)\right) \\
& \propto \exp\left(-\frac{1}{2}\left(\frac{\beta^2-2\beta\left(\frac{1}{n}\left(\sum_{i=1}^{n}z_i\left(x_i-\alpha\right)\right)\frac{n\tau_\beta^2}{\tau_\beta^2\sum_{i=1}^{n}z_i^2+\sigma^2}+\theta_\beta\frac{\sigma^2}{\tau_\beta^2\sum_{i=1}^{n}z_i^2+\sigma^2}\right)}{\frac{\sigma_\beta^2\sigma^2}{\tau_\beta^2\sum_{i=1}^{n}z_i^2+\sigma^2}}\right)\right) \\
& \propto \frac{1}{\sqrt{2\pi\left(\frac{\tau_\beta^2\sigma^2}{\tau_\beta^2\sum_{i=1}^{n}z_i^2+\sigma^2}\right)}} \\
& \quad \times \exp\left(-\frac{1}{2}\left(\frac{\left(\beta-\left(\frac{1}{n}\left(\sum_{i=1}^{n}z_i\left(x_i-\alpha\right)\right)\frac{n\tau_\beta^2}{\tau_\beta^2\sum_{i=1}^{n}z_i^2+\sigma^2}+\theta_\beta\frac{\sigma^2}{\tau_\beta^2\sum_{i=1}^{n}z_i^2+\sigma^2}\right)\right)^2}{\frac{\tau_\alpha^2\sigma^2}{\tau_\beta^2\sum_{i=1}^{n}z_i^2+\sigma^2}}\right)\right)
\end{aligned}
$$

which is a normal distribution with mean parameter

$$
\frac{1}{n}\left(\sum_{i=1}^{n}z_i\left(x_i-\alpha\right)\right)\frac{n\tau_\beta^2}{\tau_\beta^2\sum_{i=1}^{n}z_i^2+\sigma^2}+\theta_\beta\frac{\sigma^2}{\tau_\beta^2\sum_{i=1}^{n}z_i^2+\sigma^2}
$$

and variance parameter

$$
\frac{\tau_\beta^2\sigma^2}{\tau_\beta^2\sum_{i=1}^{n}z_i^2+\sigma^2}.
$$

Finally, we apply the same logic to the variance parameter, σ^2:

$$
\begin{aligned}
& f_{\Sigma^2|\mathbf{X}=\mathbf{x},\, A=\alpha,\, B=\beta}(\sigma^2) \\
&\propto \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2}\frac{\sum_{i=1}^{n}\left(x_i-\alpha-\beta z_i\right)^2}{\sigma^2}\right)\left(\frac{1}{\sigma^2}\right)^{n_\sigma/2+1}\exp\left(-\frac{\theta_\sigma/2}{\sigma^2}\right) \\
&\propto \exp\left(-\frac{1}{2}\frac{\theta_\sigma+\sum_{i=1}^{n}\left(x_i-\alpha-\beta z_i\right)^2}{\sigma^2}\right)\left(\frac{1}{\sigma^2}\right)^{(n_\sigma+n)/2+1} \\
&\propto \frac{\left(\frac{\theta_\sigma+\sum_{i=1}^{n}(x_i-\alpha-\beta z_i)^2}{2}\right)^{(n_\sigma+n)/2}}{\Gamma((n_\sigma+n)/2)}\left(\frac{1}{\sigma^2}\right)^{(n_\sigma+n)/2+1}\exp\left(-\frac{\frac{\theta_\sigma+\sum_{i=1}^{n}(x_i-\alpha-\beta z_i)^2}{2}}{\sigma^2}\right),
\end{aligned}
$$

which is an inverse gamma distribution with shape parameter $\frac{n_\sigma+n}{2}$ and scale parameter

$$\frac{\theta_\sigma+\sum_{i=1}^{n}\left(x_i-\alpha-\beta z_i\right)^2}{2}.$$

We now apply the Gibbs sampler on *real* data. The example will use motorcycle insurance data from Wasa, a Swedish insurance company, taken from `dataOhlsson` of the R package `insuranceData`; see Wolny-Dominiak and Trzesiok (2014) for more details.

```
library("insuranceData")
data(dataOhlsson)
```

This dataset contains information about the number of motorcycle accidents, their claim cost, and some risk factors (e.g., the age of the driver, the age of the vehicle, the geographic zone).

Example 8.4.1. Bayesian Linear Regression. You wish to understand the relationship between the age of the driver and the (logarithm of the) claim cost. Let x_i be the logarithm of the i^{th} claim cost and z_i be the age associated with the i^{th} claim. Further assume the following linear relationship between the two quantities:

$$x_i = \alpha + \beta z_i + \varepsilon_i,$$

where ε_i is normally distributed with mean zero and variance σ^2. Find the posterior density of the three parameters α, β, and σ^2 using the Gibbs sampler.

Solution. Let us begin by visualizing the data. Figure 1.11 reports the logarithm of the claim cost as a function of the driver's age. At first sight, it seems

that the relationship between the claim cost and age is negative, so we should expect a negative β, generally speaking.

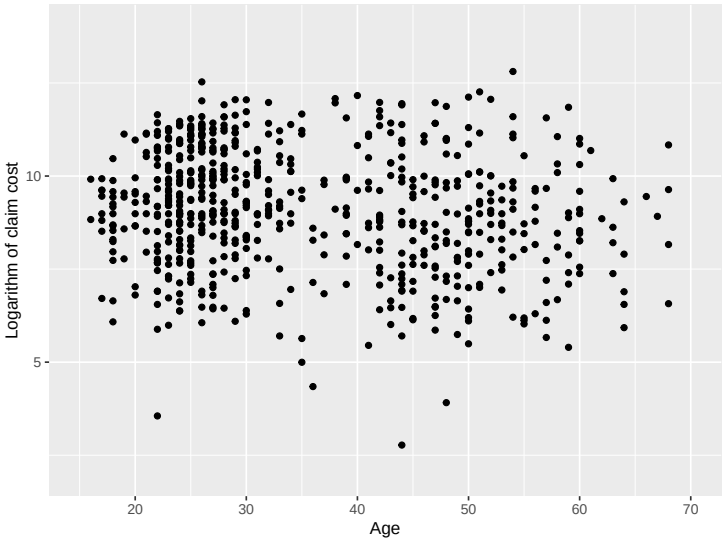

Figure 1.11: **Logarithm of the claim cost as a function of the driver's age**

Let us now turn to Bayesian inference via Gibbs sampling to find the posterior distribution of the three parameters of interest. We will use 10,000 iterations and discard the first 5,000 iterations (i.e., burn-in period). For our prior distributions, we use weakly informative priors by setting $\theta_\alpha = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$, $\theta_\beta = 0$, $\tau_\alpha^2 = \tau_\beta^2 = 10$, $n_\sigma = 1$, and $\theta_\sigma = 0.1$. The initial values of the parameters are set to: $\alpha^{(0)} = \overline{x}$, $\beta^{(0)} = 0$, and $\sigma^{2\,(0)} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$.

```
set.seed(1)
library("invgamma")
dataOhlsson <- dataOhlsson[dataOhlsson$skadkost > 0, ]
dataOhlsson$logskadkost <- log(dataOhlsson$skadkost)

x <- dataOhlsson$logskadkost
z <- dataOhlsson$agarald

n <- length(x)
M <- 10000
Mstar <- 5000
thetaa <- mean(x)
tau2a <- 10
thetab <- 0
tau2b <- 10
```

```
nsigma <- 1
thetasigma <- 0.1

alphas <- rep(NA, M + 1)
betas <- rep(NA, M + 1)
sigma2s <- rep(NA, M + 1)

alphas[1] <- mean(x)
betas[1] <- 0
sigma2s[1] <- var(x)

for (m in 2:(M + 1)) {
    # Generate alpha
    den_alpha <- n * tau2a + sigma2s[m - 1]
    mean_alpha <- (1/n) * (sum(x - betas[m - 1] * z)) * (n * tau2a)/den_alpha + thetaa
        sigma2s[m - 1]/den_alpha
    var_alpha <- tau2a * sigma2s[m - 1]/den_alpha

    alphas[m] <- rnorm(1, mean = mean_alpha, sd = sqrt(var_alpha))

    # Generate beta
    den_beta <- tau2b * sum(z^2) + sigma2s[m - 1]
    mean_beta <- (1/n) * (sum(z * (x - alphas[m]))) * (n * tau2b)/den_beta + thetab *
        sigma2s[m - 1]/den_beta
    var_beta <- tau2b * sigma2s[m - 1]/den_beta

    betas[m] <- rnorm(1, mean = mean_beta, sd = sqrt(var_beta))

    # Generate sigma
    shape_sigma <- (nsigma + n)/2
    scale_sigma <- (thetasigma + sum((x - alphas[m] - betas[m] * z)^2))/2

    sigma2s[m] <- rinvgamma(1, shape = shape_sigma, scale = scale_sigma)
}
```

Once we have the posterior parameter samples, we can get multiple quantities of interest. For instance, the posterior mean of parameters α, β, and σ^2 are 9.843, -0.0208, and 3.517×10^{-6}, respectively.

```
The posterior mean for coefficient alpha is 9.842986

The posterior mean for coefficient beta is -0.02080261

The posterior mean for the variance parameter is 3.516722e-06
```

We can also get histograms of the posterior distribution for α, β, and σ^2; Figure 1.12 reports histograms for the three parameters. The uncertainty around each

parameter is very small.

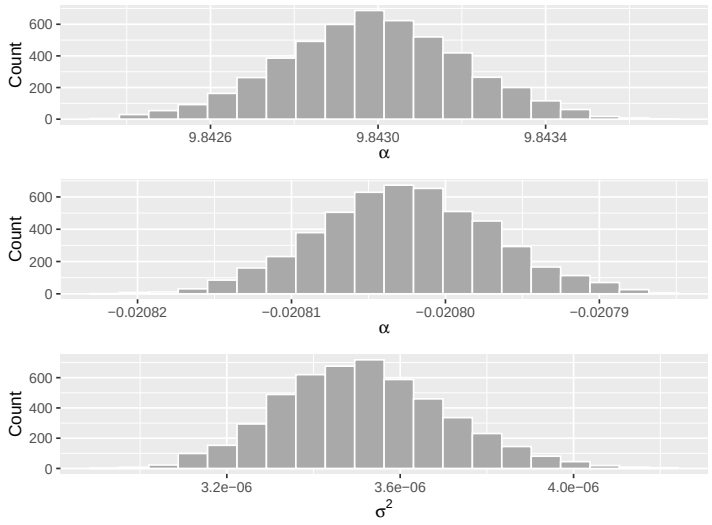

Figure 1.12: **Histogram of the posterior distribution for parameters α (top panel), β (middle panel), and σ^2 (bottom panel)**

The top panel of Figure 1.13 reports a plot of the post-burn-in values of α as a function of the iteration number; this type of plot is known as a trace plot in the literature. These samples are not impacted by the initial parameter value that was selected. Indeed, after about 20–30 iterations, the posterior parameter values obtained by the Gibbs sampler are very close to their posterior means. For instance, the bottom panel of Figure 1.13 shows a plot of the first 50 values of α as a function of the iteration number.

1.4.3 The Metropolis–Hastings Algorithm

Gibbs sampling works well when the full conditional distribution for each parameter in the model can be found and is of a common form. This, unfortunately, is not always possible, meaning that we need to rely on other computational tools to find the posterior distribution of the parameters. One very popular method that copes with the shortcomings of Gibbs' method is the Metropolis–Hastings sampler.

Let us assume that the current value of the first model parameter is $\theta_1^{(0)}$. From this current value, we now wish to find a new value for this parameter. To do so, we propose a new value for this parameter, θ_1^*, from a candidate (or proposal) density $q\left(\theta_1^* \middle| \theta_1^{(0)}\right)$. Since this proposal has nothing to do with the posterior

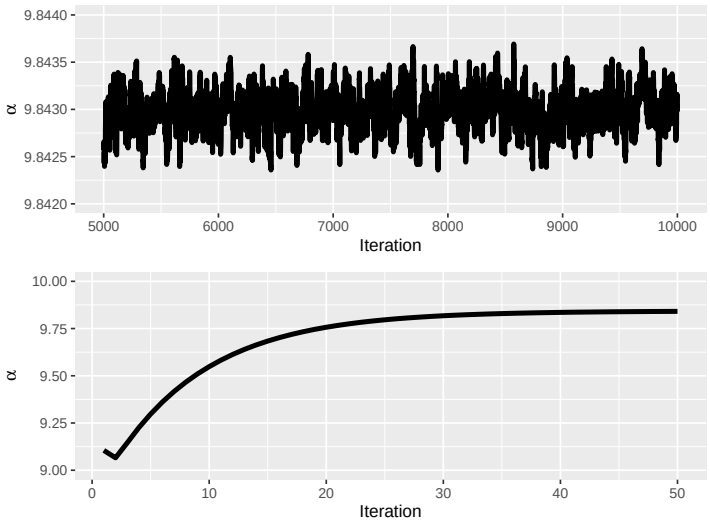

Figure 1.13: **Trace plot of α for the post-burn-in iterations (top panel) and for the first 50 iterations (bottom panel)**

distribution of the parameter, we should not keep all candidates in our final sample—we only accept those samples that are representative of the posterior distribution of interest. To determine whether we accept or reject the candidate, we compute a so-called acceptance ratio $\alpha\left(\theta_1^{(0)}, \theta_1^*\right)$ using

$$\alpha\left(\theta_1^{(0)}, \theta_1^*\right) = \frac{h\left(\theta_1^*\right) q\left(\theta_1^{(0)} \,\middle|\, \theta_1^*\right)}{h\left(\theta_1^{(1)}\right) q\left(\theta_1^* \,\middle|\, \theta_1^{(0)}\right)}$$

where

$$h(\theta_1) = f_{\mathbf{X} \,|\, \Theta_1 = \theta_1, \boldsymbol{\Theta}_{\setminus 1} = \boldsymbol{\theta}_{\setminus 1}}(\mathbf{x})\, f_{\Theta_1, \boldsymbol{\Theta}_{\setminus 1}}\left(\theta_1, \boldsymbol{\theta}_{\setminus 1}\right)$$

and $\boldsymbol{\theta}_{\setminus 1}$ represents all parameters except for the first one. Then, we accept the proposed value θ_1^* with probability $\alpha\left(\theta_1^{(0)}, \theta_1^*\right)$ and reject it with probability $1 - \alpha\left(\theta_1^{(0)}, \theta_1^*\right)$. Specifically,

$$\theta_1^{(1)} = \begin{cases} \theta_1^* & \text{with probability } \alpha\left(\theta_1^{(0)}, \theta_1^*\right) \\ \theta_1^{(0)} & \text{with probability } 1 - \alpha\left(\theta_1^{(0)}, \theta_1^*\right) \end{cases}$$

Then, we can repeat the same process for all other parameters to obtain $\theta_2^{(1)}$ to $\theta_k^{(1)}$, while replacing the parameters $\theta_{\setminus i}$ by their most current values in the

chain. Once we have updated all values, we can repeat this process for all m in $\{2, 3, ..., M\}$, similar to the iterative process used in the Gibbs sampler.

Special Case: Symmetric Proposal Distribution. If a proposal distribution is symmetric, then

$$q\left(\theta_i^{(m)} \,\middle|\, \theta_i^*\right) = q\left(\theta_i^* \,\middle|\, \theta_i^{(m)}\right),$$

and those terms cancel out, leaving

$$\alpha\left(\theta_i^{(m)}, \theta_1^*\right) = \frac{h\left(\theta_i^*\right)}{h\left(\theta_i^{(m)}\right)}.$$

This special case is called the Metropolis algorithm.

The Metropolis–Hastings sampler requires a lot of fine-tuning, generally speaking, because the experimenter needs to select one proposal distributions for each parameter. A common approach is to assume a normal proposal distribution centered at the previous value; that is,

$$\Theta_i^* \sim \text{Normal}\left(\theta_i^{(m-1)}, \sigma_i^2\right),$$

at step m, where σ_i^2 is the variance of the i^{th} parameter's proposal distribution.

Example 8.4.2. Impact of Proposal Density on the Acceptance Rate. Assume that each policyholder's claim count (frequency) is distributed as a Poisson random variable such that

$$p_{N_i \,|\, \Lambda=\lambda}(n_i) = \frac{\lambda^{n_i} e^{-\lambda}}{n_i!},$$

where n_i is the number of claims associated with the i^{th} policyholder. Further assume a noninformative, flat prior over $[0, \infty]$; that is,

$$f_\Lambda(\lambda) \propto 1, \quad \lambda \in [0, \infty].$$

Find the posterior distribution of the parameter using 1,000 iterations of the Metropolis–Hastings sampler assuming the claim count data of the Singapore Insurance Data (see Example 8.1.4 for more details). Use a normal proposal with small (1×10^{-7}), moderate (1×10^{-4}), and large (1×10^{-1}) values as the proposal variance in your tests and comment on the differences.

Solution. Starting from the the likelihood function and the prior distribution, we have that

$$h(\lambda) \propto \prod_{i=1}^{N} \frac{\lambda^{n_i} e^{-\lambda}}{n_i!}.$$

```
M <- 1000

# First variance: 1 x 10^-7
set.seed(1)
sigma21 <- 1e-07

lambdas1 <- rep(NA, M + 1)
lambdas1[1] <- mean(sgautonb$Clm_Count)
accept1 <- rep(NA, M)

for (m in 2:(M + 1)) {
    # Compute logarithm of h for past value
    loghpast <- sum(dpois(sgautonb$Clm_Count, lambda = lambdas1[m - 1], log = TRUE))

    # Generate proposed parameter and compute logarithm of h for proposed value
    lambdastar <- rnorm(1, mean = lambdas1[m - 1], sd = sqrt(sigma21))
    if (lambdastar > 0) {
        loghstar <- sum(dpois(sgautonb$Clm_Count, lambda = lambdastar, log = TRUE))
    } else {
        loghstar <- -Inf
    }

    # Compute acceptance probability and copy new parameter value
    alpha <- exp(loghstar - loghpast)
    if (alpha > runif(1, min = 0, max = 1)) {
        lambdas1[m] <- lambdastar
        accept1[m - 1] <- 1
    } else {
        lambdas1[m] <- lambdas1[m - 1]
        accept1[m - 1] <- 0
    }
}

# Second variance: 1 x 10^-4
set.seed(1)
sigma22 <- 1e-04

lambdas2 <- rep(NA, M + 1)
lambdas2[1] <- mean(sgautonb$Clm_Count)
```

```
accept2 <- rep(NA, M)

for (m in 2:(M + 1)) {
    # Compute logarithm of h for past value
    loghpast <- sum(dpois(sgautonb$Clm_Count, lambda = lambdas2[m - 1], log = TRUE))

    # Generate proposed parameter and compute logarithm of h for proposed value
    lambdastar <- rnorm(1, mean = lambdas2[m - 1], sd = sqrt(sigma22))
    if (lambdastar > 0) {
        loghstar <- sum(dpois(sgautonb$Clm_Count, lambda = lambdastar, log = TRUE))
    } else {
        loghstar <- -Inf
    }

    # Compute acceptance probability and copy new parameter value
    alpha <- exp(loghstar - loghpast)
    if (alpha > runif(1, min = 0, max = 1)) {
        lambdas2[m] <- lambdastar
        accept2[m - 1] <- 1
    } else {
        lambdas2[m] <- lambdas2[m - 1]
        accept2[m - 1] <- 0
    }
}

# Third variance: 1 x 10^-1
set.seed(1)
sigma23 <- 0.1

lambdas3 <- rep(NA, M + 1)
lambdas3[1] <- mean(sgautonb$Clm_Count)
accept3 <- rep(NA, M)

for (m in 2:(M + 1)) {
    # Compute logarithm of h for past value
    loghpast <- sum(dpois(sgautonb$Clm_Count, lambda = lambdas3[m - 1], log = TRUE))

    # Generate proposed parameter and compute logarithm of h for proposed value
    lambdastar <- rnorm(1, mean = lambdas3[m - 1], sd = sqrt(sigma23))
    if (lambdastar > 0) {
        loghstar <- sum(dpois(sgautonb$Clm_Count, lambda = lambdastar, log = TRUE))
    } else {
        loghstar <- -Inf
    }
```

```
    # Compute acceptance probability and copy new parameter value
    alpha <- exp(loghstar - loghpast)
    if (alpha > runif(1, min = 0, max = 1)) {
        lambdas3[m] <- lambdastar
        accept3[m - 1] <- 1
    } else {
        lambdas3[m] <- lambdas3[m - 1]
        accept3[m - 1] <- 0
    }
}
```

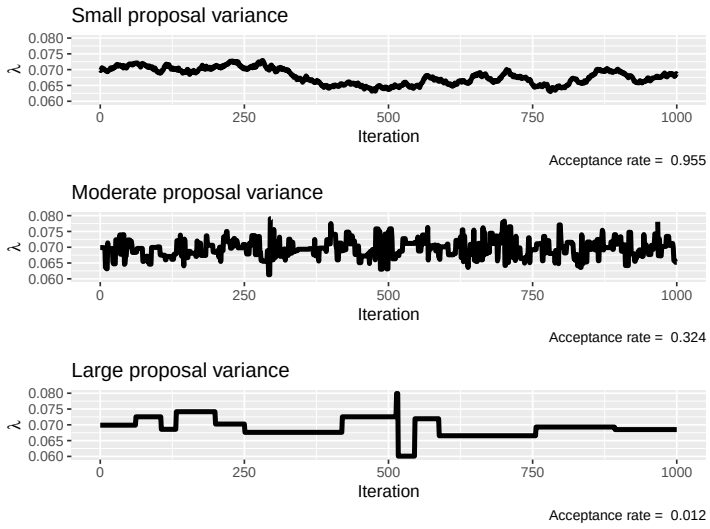

Figure 1.14: **Trace plots based on three different proposals: $\sigma^2 = 1 \times 10^{-7}$ (top panel), $\sigma^2 = 1 \times 10^{-4}$ (middle panel), and $\sigma^2 = 1 \times 10^{-1}$ (bottom panel)**

Different variance parameters lead to different results. In this example, if σ^2 is too small, then the experimenter tends to draw samples that are very similar from one iteration to the other. This increases the acceptance rate (i.e., the rate at which we accept the proposal), but also means that the chain is travelling slowly around the posterior distribution. This ultimately imply that it will take longer chains to visit the whole posterior distribution. One way to see this issue in practice is by computing autocorrelation coefficients for the sample of parameter (more details on this in Section 1.4.4). The top panel of Figure 1.14 indeed shows this strong autocorrelation and slow travelling around the posterior distribution.

On the other hand, if σ^2 is too large, then the proposal are seldom accepted, and

the chain will tend to stick—exhibiting long period for which the chain stays constant. For instance, the case with large proposal variance above leads to an acceptance rate of 1.2%, which is very low. The bottom panel of Figure 1.14 reports this issue.

The moderate proposal variance case reports an acceptance rate of 32.4%, which is not too high nor too low. The general behavior of this chain resembles that of a hairy caterpillar—a good sign—meaning that the mixing seems adequate and that we accept a decent amount of proposed values.

Finding the right proposal variance values for problems of interest requires some fine-tuning. As a general guideline, experimenters should target acceptance rates between 20% and 50%.

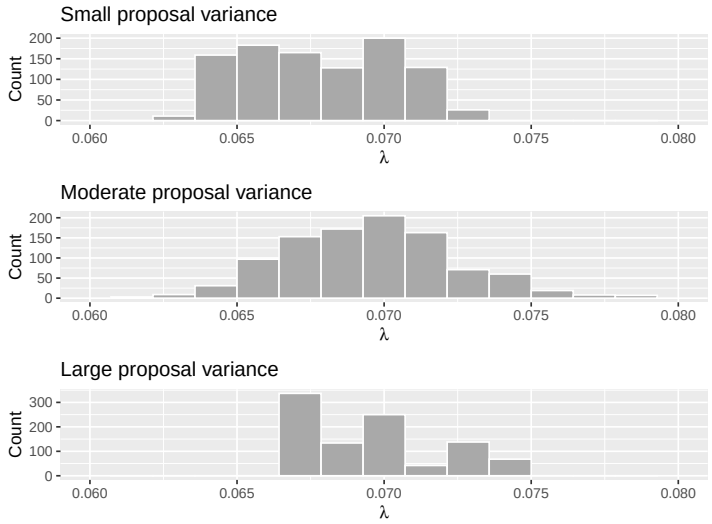

Figure 1.15: **Posterior densities based on three different proposals: $\sigma^2 = 1 \times 10^{-7}$ (top panel), $\sigma^2 = 1 \times 10^{-4}$ (middle panel), and $\sigma^2 = 1 \times 10^{-1}$ (bottom panel)**

Using the wrong proposal distribution can have an impact on the posterior distribution, as shown in Figure 1.15. A small variance takes a long time to travel throughout the posterior distribution, whereas a large variance tends to stick.

Example 8.4.3. Impact of Initial Parameters. Consider the motorcycle insurance data from Wasa used in Example 8.4.1. We wish to model the claim amount from motorcycle losses with a gamma distribution; that is,

$$f_{X_i \,|\, \Theta=\theta,\, A=\alpha}(x_i) = \frac{1}{\theta^{\alpha}\Gamma(\alpha)} x_i^{\alpha-1} e^{-\frac{x_i}{\theta}},$$

where x_i is the i^{th} claim amount. We assume that the prior distributions for both θ and α are noninformative and flat; that is,

$$f_{\Theta,A}(\theta,\alpha) \propto 1, \quad \theta \in [0,\infty], \quad \alpha \in [0,\infty].$$

Find the posterior distribution of the parameter using 1,000 iterations of the Metropolis–Hastings sampler. Use a normal proposal with a proposal variance 5×10^7 for θ and 1×10^{-2} for α, and rely on $\theta^{(0)} = 50,000$ and $\alpha^{(0)} = 0.5$ to start the Metropolis–Hastings sampler. Redo the experiment with $\theta^{(0)} = 10,000$ and $\alpha^{(0)} = 2.5$.

Solution. Starting from the the likelihood function and the prior distribution, we have that

$$h(\theta,\alpha) \propto \prod_{i=1}^{N} \frac{1}{\theta^{\alpha}\Gamma(\alpha)} x_i^{\alpha-1} e^{-\frac{x_i}{\theta}}.$$

```
dataOhlsson <- dataOhlsson[dataOhlsson$skadkost > 0, ]
x <- dataOhlsson$skadkost
M <- 1000

set.seed(1)
sigma2theta <- 5e+07
sigma2alpha <- 0.01

# First set of initial values
thetas1 <- rep(NA, M + 1)
thetas1[1] <- 50000
alphas1 <- rep(NA, M + 1)
alphas1[1] <- 0.5

acceptthetа1 <- rep(NA, M)
acceptalpha1 <- rep(NA, M)

for (m in 2:(M + 1)) {
    # Let us start with theta Compute logarithm of h for past value
    loghpast <- sum(dgamma(x, scale = thetas1[m - 1], shape = alphas1[m - 1], log = TRU

    # Generate proposed parameter and compute logarithm of h for proposed value
    thetastar <- rnorm(1, mean = thetas1[m - 1], sd = sqrt(sigma2theta))
    if (thetastar > 0) {
```

```
        loghstar <- sum(dgamma(x, scale = thetastar, shape = alphas1[m - 1], log = TRUE))
    } else {
        loghstar <- -Inf
    }

    # Compute acceptance probability and copy new parameter value
    alphatheta <- exp(loghstar - loghpast)
    if (alphatheta > runif(1, min = 0, max = 1)) {
        thetas1[m] <- thetastar
        acceptttheta1[m - 1] <- 1
    } else {
        thetas1[m] <- thetas1[m - 1]
        acceptttheta1[m - 1] <- 0
    }

    # And then deal with alpha Compute logarithm of h for past value
    loghpast <- sum(dgamma(x, scale = thetas1[m], shape = alphas1[m - 1], log = TRUE))

    # Generate proposed parameter and compute logarithm of h for proposed value
    alphastar <- rnorm(1, mean = alphas1[m - 1], sd = sqrt(sigma2alpha))
    if (thetastar > 0) {
        loghstar <- sum(dgamma(x, scale = thetas1[m], shape = alphastar, log = TRUE))
    } else {
        loghstar <- -Inf
    }

    # Compute acceptance probability and copy new parameter value
    alphaalpha <- exp(loghstar - loghpast)
    if (alphaalpha > runif(1, min = 0, max = 1)) {
        alphas1[m] <- alphastar
        acceptalpha1[m - 1] <- 1
    } else {
        alphas1[m] <- alphas1[m - 1]
        acceptalpha1[m - 1] <- 0
    }
}

# Second set of initial values
set.seed(1)
thetas2 <- rep(NA, M + 1)
thetas2[1] <- 10000
alphas2 <- rep(NA, M + 1)
alphas2[1] <- 2.5

acceptttheta2 <- rep(NA, M)
```

```
acceptalpha2 <- rep(NA, M)

for (m in 2:(M + 1)) {
    # Let us start with theta Compute logarithm of h for past value
    loghpast <- sum(dgamma(x, scale = thetas2[m - 1], shape = alphas2[m - 1], log = TRU

    # Generate proposed parameter and compute logarithm of h for proposed value
    thetastar <- rnorm(1, mean = thetas2[m - 1], sd = sqrt(sigma2theta))
    if (thetastar > 0) {
        loghstar <- sum(dgamma(x, scale = thetastar, shape = alphas2[m - 1], log = TRU
    } else {
        loghstar <- -Inf
    }

    # Compute acceptance probability and copy new parameter value
    alphatheta <- exp(loghstar - loghpast)
    if (alphatheta > runif(1, min = 0, max = 1)) {
        thetas2[m] <- thetastar
        acceptttheta2[m - 1] <- 1
    } else {
        thetas2[m] <- thetas2[m - 1]
        acceptttheta2[m - 1] <- 0
    }

    # And then deal with alpha Compute logarithm of h for past value
    loghpast <- sum(dgamma(x, scale = thetas1[m], shape = alphas2[m - 1], log = TRUE))

    # Generate proposed parameter and compute logarithm of h for proposed value
    alphastar <- rnorm(1, mean = alphas2[m - 1], sd = sqrt(sigma2alpha))
    if (thetastar > 0) {
        loghstar <- sum(dgamma(x, scale = thetas2[m], shape = alphastar, log = TRUE))
    } else {
        loghstar <- -Inf
    }

    # Compute acceptance probability and copy new parameter value
    alphaalpha <- exp(loghstar - loghpast)
    if (alphaalpha > runif(1, min = 0, max = 1)) {
        alphas2[m] <- alphastar
        acceptalpha2[m - 1] <- 1
    } else {
        alphas2[m] <- alphas2[m - 1]
        acceptalpha2[m - 1] <- 0
    }
}
```

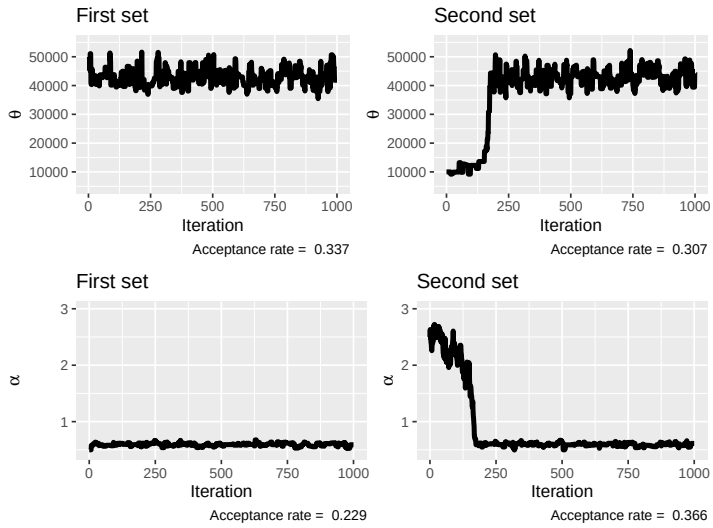

Figure 1.16: **Trace plots based on two different starting parameter sets: $\theta^{(0)} = 50,000$ and $\alpha^{(0)} = 0.5$ (left panels), and $\theta^{(0)} = 10,000$ and $\alpha^{(0)} = 2.5$ (right panels)**

Clearly, from Figure 1.16, the initial parameter value matters: for the first set, the starting value is close to the posterior mode, meaning that the final sample does not depend much on the starting value. For the second set, on the other hand, it takes about 200 iterations to get closer to where most of the density resides. Having a burn-in in the case of Metropolis–Hastings sampler is therefore a good idea to reduce the impact of initial guesses on the final posterior distribution.

In the next subsection, we learn a few methods and metrics to diagnose the convergence of the Markov chains generated via MCMC methods.

1.4.4 Markov Chain Diagnostics

There are many different tuning parameters in MCMC schemes, and they all have an impact on the convergence of the Markov chains generated by these methods. To understand the impact of these choices on the chains (e.g., number of iterations, length of burn-in, proposal distribution), we introduce a few methods to analyze their convergence.

Examining Trace Plots and Autocorrelation

Trace Plot. The most elementary tool to assess whether MCMC chains have converged to the posterior distribution is the trace plot. As mentioned above, a trace plot displays the sequence of samples as a function of the iteration number, with the sample value on the y-axis and the iteration number on the x-axis. If the chain has converged, the trace plot should show a stable sequence of samples around the true posterior distribution that looks like a hairy caterpillar. However, if the chain has not yet converged, the trace plot may show a sequence of samples that still appear to be changing or have not yet settled into a stable pattern.

In addition to assessing convergence, trace plots can also be used to diagnose potential problems with MCMC algorithms, such as poor mixing or autocorrelation. For example, if the trace plot shows long periods of no change followed by abrupt jumps, this may indicate poor mixing and suggest that the MCMC algorithm needs to be adjusted or a different method should be used.

Lag-1 Autocorrelation. Another quantity that might be helpful is the lag-1 autocorrelation—the correlation between consecutive samples in a given chain:

$$\operatorname{Cov}\left[\theta_i^{(m)}, \theta_i^{(m-1)}\right].$$

Note that if the autocorrelation is too high, it can indicate that the chain is not mixing well and is not sampling the posterior distribution effectively. This can result in poor convergence, longer run times, and decreased precision of the estimates obtained from the MCMC algorithm.

In addition to examining trace plots and computing autocorrelation coefficients, we can use other, more formal tools to evaluate whether the chains obtained are reliable and have converged.

Comparing Parallel Chains

Gelman–Rubin Statistic Another way to assess convergence is to run multiple chains in parallel from different starting points and check if their behavior is similar. In addition to comparing their trace plots, the chains can be compared by using a statistical test—the Gelman–Rubin test of Gelman and Rubin (1992). The latter test compares the within-chain variance to the between-chain variance; to calculate the statistic, we need to generate a small number of chains (say, R), each for $M - M^*$ post-burn-in iterations.

If the chains have converged, the within-chain variance should be similar to the between-chain variance. Assuming the parameter of interest is θ_i, the within-chain variance is

$$W = \frac{1}{R(M - M^* - 1)} \sum_{r=1}^{R} \sum_{m=M^*+1}^{M} \left(\theta_{i,r}^{(m)} - \overline{\theta}_{i,r}\right)^2,$$

where $\theta_{i,r}^{(m)}$ is the m^{th} draw of θ_i in the r^{th} chain and $\overline{\theta}_{i,r}$ is the sample mean of θ_i for the r^{th} chain. The between-chain variance is given by

$$B = \frac{M - M^*}{R - 1} \sum_{r=1}^{R} \left(\overline{\theta}_{i,r} - \overline{\theta}_i\right),$$

where $\overline{\theta}_i$ is the overall sample mean of θ_i from all chains. The Gelman–Rubin statistic is

$$\sqrt{\left(\frac{M - M^* - 1}{M - M^*} + \frac{R + 1}{R(M - M^*)} \frac{B}{W}\right) \frac{\text{df}}{\text{df} - 2}},$$

where df is the degrees of freedom from Student's t-distribution that approximates the posterior distribution. The statistic should produce a value close to 1 if the chain has converged. On the other hand, if the statistic value is greater than 1.1 or 1.2, this indicates that the chains may not have converged, and further analysis may be needed to determine why the chains are not mixing well.

Calculating Effective Sample Sizes

Effective Sample Size The effective sample size (ESS) is a measure of the number of independent samples obtained from an MCMC chain. Recall that in an MCMC chain, each sample is correlated with the previous sample; as a result, the effective number of independent samples is usually much smaller than the total number of samples generated by the MCMC algorithm. The ESS takes this correlation into account and provides an estimate of the number of independent samples that are equivalent to the correlated samples in the chain.

In general, a higher effective sample size indicates that the MCMC algorithm has produced more independent samples and is more likely to have accurately sampled the posterior distribution. A lower effective sample size, on the other hand, suggests that the MCMC algorithm may require further tuning or optimization to produce reliable posterior estimates.

The function `multiESS`of the R package `mcmcse` contains a function that gives the ESS of a multivariate Markov chain as described in Vats et al. (2019). The package also includes an estimate of the minimum ESS required for a specified relative tolerance level (see function `minESS`).

We now apply these various diagnostics to an example.

Example 8.4.4. Markov Chain Diagnostics. Consider the setup of Example 8.4.2. Using chains of 51,000 iterations and a burn-in of 1,000 iterations, calculate the various Markov chain diagnostics mentioned above.

Solution. Let us begin by generating five chains.

```
M <- 51000
Mstar <- 1000
R <- 5

set.seed(1)
sigma2 <- 1e-04

lambdas <- matrix(data = NA, ncol = M + 1, nrow = R)
accept <- matrix(data = NA, ncol = M, nrow = R)

for (r in 1:R) {
    lambdas[r, 1] <- max(1e-06, mean(sgautonb$Clm_Count) + rnorm(1, mean = 0, sd = 0.1)

    for (m in 2:(M + 1)) {
        # Compute logarithm of h for past value
        loghpast <- sum(dpois(sgautonb$Clm_Count, lambda = lambdas[r, m - 1], log = TRU

        # Generate proposed parameter and compute logarithm of h for proposed
        # value
        lambdastar <- rnorm(1, mean = lambdas[r, m - 1], sd = sqrt(sigma2))
        if (lambdastar > 0) {
            loghstar <- sum(dpois(sgautonb$Clm_Count, lambda = lambdastar, log = TRUE))
        } else {
            loghstar <- -Inf
        }

        # Compute acceptance probability and copy new parameter value
        alpha <- exp(loghstar - loghpast)
        if (alpha > runif(1, min = 0, max = 1)) {
            lambdas[r, m] <- lambdastar
            accept[r, m - 1] <- 1
        } else {
            lambdas[r, m] <- lambdas[r, m - 1]
            accept[r, m - 1] <- 0
        }
    }
}
save(lambdas, accept, sgautonb, alpha, lambdastar, loghpast, loghstar, m, M, Mstar,
    r, R, sigma2, file = "../IntermediateCalcs/BayesChap/Example844.Rdata")
```

Figure 1.17 reports the trace plot for the first chain: it indeed looks like a hairy caterpillar, which is a good sign.

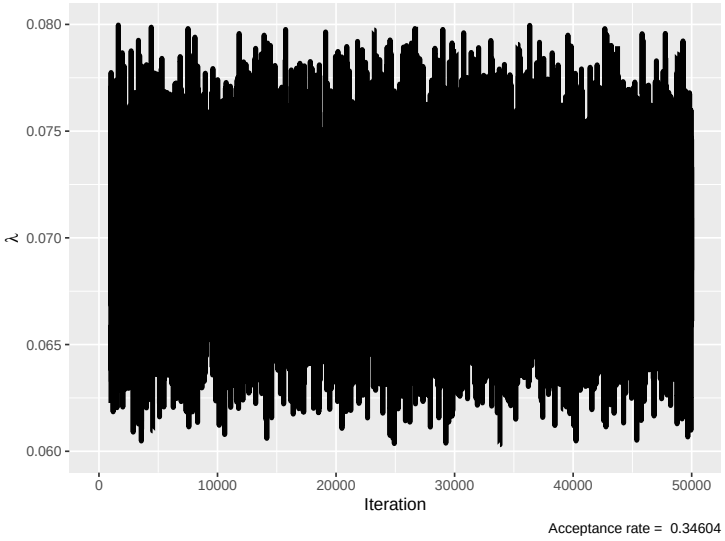

Figure 1.17: **Trace plot for parameter λ**

```
autocorr <- acf(lambdas[1, (Mstar + 1):M], lag.max = 1, plot = FALSE)
cat("The lag-1 autocorrelation coefficient is", autocorr$acf[2])
```

```
The lag-1 autocorrelation coefficient is 0.6515109
```

The autocorrelation is also mild at 65%, again pointing towards good convergence behavior.

```
W <- 0
B <- 0
for (r in 1:R) {
    W <- W + 1/(R * (M - Mstar - 1)) * sum((lambdas[r, (Mstar + 1):M] - mean(lambdas[r,
        (Mstar + 1):M]))^2)
    B <- B + (M - Mstar)/(R - 1) * sum((mean(lambdas[r, (Mstar + 1):M]) - mean(lambdas[,
        (Mstar + 1):M]))^2)
}
# Assuming that df/(df-2) tends to 1
GR <- sqrt((M - Mstar - 1)/(M - Mstar) + (R + 1)/(R * (M - Mstar)) * B/W)
cat("The Gelman-Rubin statistic is", GR)
```

```
The Gelman-Rubin statistic is 1.000127
```

The Gelman–Rubin statistic is very close to 1 in this case, meaning that the chain converged.

```
library("mcmcse")
ess <- multiESS(matrix(lambdas[1, ]))
mess <- minESS(p = 1)
cat("The ESS is", ess, "and the minimum ESS is", mess)
```

```
The ESS is 9927.299 and the minimum ESS is 6146
```

The last diagnostic refers to the ESS, and its comparison to the minimum ESS. In our case, the ESS is about 9,927, and the minimum ESS is 6,146. Since our ESS is above the minimum, we know we have a large enough sample to adequately capture the posterior distribution of λ.

1.5 Further Resources and Contributors

Many great books exist on Bayesian statistics and MCMC schemes. We refer the interested reader to Bernardo and Smith (2009) and Robert and Casella (1999) for an advanced treatment of these topics.

Contributors

- **Jean-François Bégin**, Simon Fraser University, is the principal author of the initial version of this chapter. Email: jbegin@sfu.ca for chapter comments and suggested improvements.
- Chapter reviewers include: Brian Hartman.

Bibliography

Bernardo, José M and Adrian FM Smith (2009). *Bayesian Theory*. John Wiley & Sons: New York, NY, United States of America.

Cowles, Mary Kathryn (2013). *Applied Bayesian Statistics: With R and OpenBUGS Examples*. Springer Science & Business Media: New York, NY, United States of America.

Gelfand, Alan E and Adrian FM Smith (1990). "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, Vol. 85, pp. 398–409.

Gelman, Andrew and Donald B Rubin (1992). "Inference from iterative simulation using multiple sequences," *Statistical Science*, pp. 457–472.

Hastings, WK (1970). "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, Vol. 57, pp. 97–109.

Metropolis, Nicholas, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller (1953). "Equation of state calculations by fast computing machines," *Journal of Chemical Physics*, Vol. 21, pp. 1087–1092.

Meyers, Glenn (1994). "Quantifying the Uncertainty in Claim Severity Estimates for an Excess Layer When Using the Single Parameter Pareto," in *Proceedings of the Casualty Actuarial Society*, Vol. 81, pp. 91–122.

O'Donnell, Terence (1936). *History of Life Insurance in its Formative Years*. American Conservation Company: Chicago, IL, United States of America.

Robert, Christian P and George Casella (1999). *Monte Carlo Statistical Methods*. Springer: New York, NY, United States of America.

Vats, Dootika, James M Flegal, and Galin L Jones (2019). "Multivariate output analysis for Markov chain Monte Carlo," *Biometrika*, Vol. 106, pp. 321–337.

Wolny-Dominiak, Alicja and Michal Trzesiok (2014). "Package 'insuranceData',"Technical report, The Comprehensive R Archive Network.